

N° d'ordre : 553 I

THÈSE

Présentée et soutenue par

Kafil Hajlaoui

pour obtenir le titre de

Docteur en Sciences

de l'Ecole Nationale Supérieure des Mines de Saint-Etienne
Mention : Informatique

Dispositifs de recherche et de traitement de l'information en vue d'une aide à la constitution de réseaux d'entreprises

Soutenue à Saint Etienne, le 8 décembre 2009

En présence d'un jury composé de :

Président :	Robert Mahl	Professeur	ENSM de Paris
Rapporteurs :	Robert Mahl	Professeur	ENSM de Paris
	Omar Boussaid	Professeur	Université de Lyon2
Examinateurs :	Eric Bonjour	HDR	Université de Franche-Comté
	Michel Beigbeder	Maître-Assistant	ENSM de Saint Etienne
Directeurs de thèse :	Jean Jacques Girardot	Maître de recherche	ENSM de Saint Etienne
	Xavier Boucher	HDR	ENSM de Saint Etienne

Spécialités doctorales :	Responsable :
SCIENCES ET GENIE DES MATERIAUX	J. DRIVER Directeur de recherche - Centre SMS
MECANIQUE ET INGENIERIE	A. VAUTRIN Professeur - Centre SMS
GENIE DES PROCEDES	G. THOMAS Professeur - Centre SPIN
SCIENCES DE LA TERRE	B. GUY Maitre de recherche
SCIENCES ET GENIE DE L'ENVIRONNEMENT	J. BOURGOIS Professeur - Centre SITE
MATHEMATIQUES APPLIQUEES	E. TOUBOUL Ingénieur- Centre G2I
INFORMATIQUE	O. BOISSIER Professeur - Centre G2I
IMAGE, VISION, SIGNAL	JC. PINOLI Professeur - Centre CIS
GENIE INDUSTRIEL	P. BURLAT Professeur - Centre G2I
MICROELECTRONIQUE	Ph. COLLOT Professeur - Centre CMP

AVRIL	Stéphane	MA	Mécanique & Ingénierie	CIS
BATTON-HUBERT	Mireille	MA	Sciences & Génie de l'Environnement	SITE
BENABEN	Patrick	PR 2	Sciences & Génie des Matériaux	CMP
BERNACHE-ASSOLANT	Didier	PR 1	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	MR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 2	Informatique	G2I
BOUCHER	Xavier	MA	Génie industriel	G2I
BOUDAREL	Marie-Reine	MA	Sciences de l'inform. & com.	DF
BOURGOIS	Jacques	PR 1	Sciences & Génie de l'Environnement	SITE
BRODHAG	Christian	MR	Sciences & Génie de l'Environnement	SITE
BURLAT	Patrick	PR 2	Génie industriel	G2I
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 1	Génie des Procédés	DF
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DARRIEULAT	Michel	ICM	Sciences & Génie des Matériaux	SMS
DECHOMETS	Roland	PR 2	Sciences & Génie de l'Environnement	SITE
DESRAYAUD	Christophe	MA	Mécanique & Ingénierie	SMS
DELAFOSSE	David	PR 2	Mécanique & Ingénierie	SMS
DOLGUI	Alexandre	PR 1	Informatique	G2I
DRAPIER	Sylvain	PR 2	Mécanique & Ingénierie	SMS
DRIVER	Julian	DR	Sciences & Génie des Matériaux	SMS
FEILLET	Dominique	PR2	Génie Industriel	CMP
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	CIS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	SITE
FORTUNIER	Roland	PR 1	Sciences & Génie des Matériaux	SMS
FRACZKIEWICZ	Anna	MR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	CR	Génie des Procédés	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	G2I
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GOEURIOT	Patrice	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	SITE
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUILHOT	Bernard	DR	Génie des Procédés	CIS
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
INAL	Karim	MR	Microélectronique	CMP
KLÖCKER	Helmut	CR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	SITE
LERICHE	Radolphe	CR	Mécanique & Ingénierie	SMS
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
LONDICHE	Henry	MR	Sciences & Génie de l'Environnement	SITE
MOLIMARD	Jérôme	MA	Sciences & Génie des Matériaux	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	MA1	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 1	Image, Vision, Signal	CIS
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	CR	Sciences & Génie de l'Environnement	DF
THOMAS	Gérard	PR 1	Génie des Procédés	SPIN
VALDIVIESO	Françoise	CR	Génie des Procédés	SPIN
VAUTRIN	Alain	PR 1	Mécanique & Ingénierie	SMS
VIRICELLE	Jean-Paul	CR	Génie des procédés	SPIN
WOLSKI	Krzysztof	CR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

Glossaire :

PR 1	Professeur 1ère catégorie
PR 2	Professeur 2ème catégorie
MA(MDC)	Maître assistant
DR 1	Directeur de recherche
Ing.	Ingénieur
MR(DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
ICM	Ingénieur en chef des mines

Centres :

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
SITE	Sciences Information et Technologies pour l'Environnement
G2I	Génie Industriel et Informatique
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

**Ecole Nationale Supérieure des Mines
de Saint-Etienne**

N^o d'ordre : 553 I

Kafil HAJLAOUI

Devices of research and data processing to help the networks constitution of enterprises

Computer Science

Retrieval Information, Extraction Information, Ontology, Company Networks

Abstract The indissociable industrial context of the evolution of Communication and Information Technologies today brings new forms of organizations strongly based on collaborations between firms. In this context of collaborative networks, the quality emergence of the new partnerships depends largely on the treatment and the share of information. Within the framework of virtual organisations, we are developing a decision support approach to assist the identification of collaborative corporate networks. This approach is based on an automated procedure of information extraction aiming to identify key features of potential partners. The added value of this research is to operate in an "open universe" of potential partners, using the company's public web sites as the main source of information. The key features we are extracting concern activity fields and competencies of the firms.

This research consists in the realisation of search systems of automatic extraction of information starting from the Web (web site of the companies). The objective is to meet the needs for an opened informational environment, concerning the companies. The thesis aims at developing targeted mechanisms of extraction of information, which will be used preliminary to the application of decision-making tools aid in the field of inter-company collaborations between firms. The contribution is based on a major semantic representation of information while being based on the semantic ontology, bonds and a linguistic treatment articulated around the use of the syntactic pattern. Two mechanisms of information extraction are installed, one directed on the identification of the sectors lines of business and the other directed on the extraction of companies' competences.

**Ecole Nationale Supérieure des Mines
de Saint-Etienne**

N^o d'ordre : 553 I

Kafil HAJLAOUI

Dispositifs de recherche et de traitement de l'information en vue d'une aide à la constitution de réseaux d'entreprises

Informatique

Recherche d'information, Extraction d'information, Ontologie, Réseaux d'entreprises

Résumé Le contexte industriel indissociable de l'évolution des Technologies de l'Information et de la Communication donne naissance aujourd'hui à de nouvelles formes d'organisations fortement basées sur les collaborations inter-entreprises. Dans ce contexte de réseaux collaboratifs, la qualité de l'émergence de nouveaux partenariats dépend largement des dispositifs de traitement et de partage de l'information. La recherche d'information pertinentes caractérisant les entreprises devient un outil indispensable aux managers et aux divers acteurs économiques, en vue de détecter des liens de collaboration potentiels. Dans le cadre de ces travaux de thèse, nous avons ciblé la complémentarité des activités et la similarité des compétences comme informations clés destinées à analyser les opportunités d'émergences de collaborations inter-entreprises.

Ce travail de recherche s'inscrit dans le cadre de la mise en œuvre de systèmes de recherche et d'extraction automatique d'information à partir du web (site web des entreprises). L'objectif est de répondre aux besoins d'un environnement informationnel ouvert, concernant les entreprises. La thèse vise à développer des mécanismes ciblés d'extraction d'information, dont l'utilisation sera préalable à l'application d'outil d'aide à la décision dans le domaine des collaborations inter-entreprises. La contribution est basée sur une représentation sémantique de l'information en se basant sur les ontologies, les liens sémantiques et un traitement linguistique articulé sur l'utilisation des patrons syntaxiques. Deux mécanismes d'extraction d'information sont mis en place, l'un orienté sur l'identification des secteurs d'activités des entreprises et l'autre sur le repérage de leurs compétences.

Remerciements

Je tiens, tout d'abord, à exprimer ma profonde gratitude à mes directeurs de thèse : Jean Jacques Girardot et Xavier Boucher. Leurs conseils, leur confiance et leurs encouragements ont largement contribué à l'aboutissement de ce travail.

Je remercie vivement Messieurs Robert MAHL, Professeur à l'école des mines de Paris et Omar BOUSAID, Professeur à l'université lumière 2 de Lyon, pour l'honneur qu'ils me font en acceptant d'être les rapporteurs de ce mémoire ; leur lecture attentive et leurs remarques ont permis d'en améliorer la rédaction.

Je suis extrêmement reconnaissant à Monsieur Eric BONJOUR, Maître de conférence à l'université de Franche-Comté et Michel Beigbeder, Maître assistant à l'école des mines de Saint Etienne, d'avoir examiner mes travaux et de participer au jury de ma thèse.

Je remercie Madame Michaela Mathieu avec qui j'ai eu le plaisir de travailler, pour son encouragement et l'intérêt qu'elle a manifesté pour cette thèse.

Mes remerciements vont également vers tous les membres du Centre Génie Industriel et Informatique que j'ai côtoyés durant ces trois années. Je les remercie pour leur accueil, leur soutient et leur convivialité. Un merci particulier à Ali Harb, Marie Line, Liliane et Gabrielle pour leur sympathie et leur gentillesse.

Je remercie également mes collègues du laboratoire ERIC de l'Université de Lyon Lumière 2 pour leur accueil chaleureux dans l'équipe pédagogique et la compréhension dont ils ont fait preuve en ce début d'année à l'emploi du temps chargé. Un merci particulier à Jacques Viallaneix et Cécile Favre pour leur aide et leur gentillesse.

Enfin, ma gratitude et mes remerciements s'adressent à ma famille qui m'a toujours encouragé et soutenu dans les moments difficiles.

Table des matières

CHAPITRE 1

Introduction

1.1 Introduction

1.1.1 Introduction Générale

L'évolution de l'économie, la concurrence, la pression des donneurs d'ordre et l'impact des nouvelles technologies de l'information et de la communication (TIC) sont quelques unes des raisons qui amènent les entreprises à envisager des collaborations techniques et économiques. La collaboration inter-entreprises intervient lorsque plusieurs entreprises décident de mettre en commun des informations, des ressources ou des compétences dans la poursuite d'objectifs conjoints, qui pourront déboucher sur des activités coordonnées voire intégrées. Par exemple, deux entreprises peuvent collaborer parce que chacune possède une partie de l'information, de l'expertise et des ressources nécessaires à la mise au point d'un produit. Cet aspect collaboratif dans les réseaux des entreprises nécessite de mettre en place différentes architectures pour la gestion des processus de collaboration et différentes méthodes et outils d'aide à la décision stratégique pour l'entreprise. Le développement d'approches de type décisionnel requiert de déployer des solutions pertinentes de traitement de l'information, qui pourront devenir le support de processus de pilotage des activités et processus ou encore de processus de pilotage des systèmes de compétences.
Dans l'environnement économique moderne, caractérisé par des mutations incessantes, les entreprises sont appelées à être adaptatives, flexibles et proactives. Pour cela, elles construisent des espaces coopératifs dans lesquels elles travaillent et réagissent ensemble. Ces espaces coopératifs, appelés le plus souvent "nouvelles formes organisationnelles", ont émergé dans les années 80 sous diverses formes (réseaux d'entreprises, entreprises virtuelles, clusters, groupements de PME...). Toute entreprise développe aujourd'hui des liens et des relations de différents types et avec divers partenaires en fonction de ses objectifs, besoins et caractéristiques. Cette multiplicité et diversité des liens a amené les dirigeants, mais aussi les chercheurs, à prendre en compte l'entreprise avec l'ensemble de ses ramifications : l'entreprise étendue.
Au plan scientifique, ce phénomène organisationnel impacte plusieurs domaines : management, sciences de gestion... Mais aussi recherche d'information. Les travaux d'études et d'analyses de la coopération ont commencé par définir le "Pourquoi" et les objectifs de la formation des nouvelles formes organisationnelles. Aujourd'hui, les travaux se focalisent plutôt sur le "Comment" gérer ses coopérations et le choix des partenaires. Les systèmes d'informations inter-organisationnels sont de plus en plus étudiés pour améliorer la gestion de la coopération inter-entreprises. Notre travail porte précisément sur le « comment »gérer cette coopération inter-entreprises, avec

une contribution concernant la mise en place de mécanismes d'extraction d'information, utilisés comme support pour le déploiement de mécanismes décisionnels.

1.1.2 Recherche d'information pour la collaboration inter-entreprises

Lors de la mise en place de collaborations inter-entreprises ou d'entreprises virtuelles, le système d'information peut apporter une valeur ajoutée particulière au processus collaboratif, en fournissant des solutions techniques permettant l'échange d'informations et de connaissances de caractère parfois confidentiel et stratégique, mais dont le partage est nécessaire au développement des relations économiques entre partenaires. Les recherches sur les systèmes d'informations collaboratifs ou encore sur les plateformes de mise en réseau d'entreprises ont apporté de nombreux éléments de réponses en ce sens [1] [124] [123]. Plus que jamais au cœur de cette coopération inter-entreprises, l'information doit être facilement accessible et immédiatement exploitable par les différents acteurs de l'entreprise : les collaborateurs. Les solutions mises en places pour les entreprises afin d'apporter une réponse pertinente aux besoins de leurs collaborateurs en matière de recherche d'information rencontrent encore de réels problèmes d'efficacité [11] :

- Perte de productivité : le temps consacré à rechercher une information constitue une perte de temps pour la réalisation d'autres tâches à plus forte valeur ajoutée.
- Perte de valeur : une information non accessible est une information qui n'apporte pas de valeur ajoutée pour l'entreprise. De nombreuses entreprises rencontrent des difficultés pour fournir à leurs collaborateurs un accès performant à leur information. Ce problème est dû à la complexité grandissante du système d'information et à la croissance importante du volume d'information.
- Risque d'erreurs : un mauvais accès à l'information peut faire remonter des informations erronées sans qu'on puisse les identifier avant leur utilisation.

A ces problèmes d'efficacité viennent s'ajouter les difficultés supplémentaires pour une gestion simple et efficace de l'accès à l'information : l'information produite par l'entreprise est généralement sous une forme non structurée.

Une étude menée par Ark Group [2] en septembre 2005 montre que dans 86% des entreprises interrogées, provenant de différents secteurs d'activités, la solution de recherche d'information est avant tout destinée à améliorer l'accès à l'information pour ces collaborateurs. Plus de la moitié (58%) déclarent utiliser leur solution de recherche d'information pour améliorer la prise de décisions. Par ailleurs 40% des entreprises interrogées emploient leur solution de recherche d'information pour avoir une meilleure vision de leur marché et de la concurrence. Ainsi, la recherche d'information est aujourd'hui considérée comme un service indispensable à l'ensemble des collaborateurs d'une entreprise. De plus la maturité

1. EDI x12 standards, http ://www.x12.org/
2. The Age of Search, Ark Group, octobre 2005

des solutions de recherche d'information leur a fait prendre conscience des possibilités et des enjeux offerts par les solutions technologiques actuellement disponibles.

Les exigences formulées précédemment sur la nécessité pour les partenaires de décrire leurs données (vue informationnelle), leurs ressources (vue des ressources) et leurs activités (vue fonctionnelle) permettent de pallier à des manques vis-à-vis de ces trois vues [156] : le manque de capacité des systèmes d'information à partager l'information avec une sémantique et une compréhension commune de point de vue de leur signification et leur interprétation dans un contexte de collaboration. Les besoins en matière d'information et de partage d'information chez les collaborateurs, et la complexité de cette information traduisent la nécessité du recours direct à des solutions de traitement de l'information trés pointues.

1.2 Problématique

1.2.1 Enjeux de la Thèse

Notre travail contribuant à la recherche d'information pour la collaboration inter-entreprises est né du constat ci-dessus des faiblesses des solutions techniques offertes aujourd'hui, notamment en ce qui concerne la richesse et la pertinence des informations que sont suceptibles de fournir les systèmes d'extraction d'information traditionnels. Ainsi l'enjeu de notre projet est de contribuer à une automatisation de la recherche de certaines informations clées caractérisant les entreprises, en vue d'appliquer ultérieurement des modèles formels d'aide à la décision qui visent à identifier des collaborations inter-entreprises.
Ce travail de recherche s'articule avec des travaux antérieurs développés dans le domaine du génie industriel, au sein de notre laboratoire[31] [15]. Ces travaux, centrés sur la collaboration inter-entreprises et les organisations virtuelles, ont proposé des méthodes et des outils d'aide à la décision pour la construction de réseaux d'entreprises. Ils ont notamment permis d'identifier les informations caractéristiques des entreprises susceptibles d'être utilisées en vue d'analyser les opportunités de collaborations économiques, dans une perspective dynamique. Ces outils sont basés sur la collecte et le traitement des données concernant les entreprises. Ces données sont collectées manuellement à partir d'un questionnaire rempli par les dirigeants d'entreprises. Il s'avère que les dirigeants ne sont pas toujours collaboratifs et actifs pour fournir l'information pertinente, ce qui represente une limite majeure pour ces outils. Notre contribution vise à repousser cette limite en proposant des méthodes automatiques de collecte et de traitement des données, s'appuyant sur la mise en œuvre de techniques informatiques pointues en matière de recherche d'informations. Ces méthodes sont développées dans un environnement ouvert, utilisant l'information publique, pour la recherche de partenaires. Elles reposent sur la recherche et l'extraction d'information à partir des sites web des entreprises.

1.2.2 Positionnement par rapport aux organisations virtuelles

Les systèmes d'aide à la décision jouent un rôle important dans le processus de la création et la gestion de la collaboration. [84] [94] soulignent le manque des méthodes et des outils d'aide à la décision dans la construction des structures coopératives. Cette aide à la décision requiert des niveaux stratégiques et tactiques de la gestion : un système de gestion coopératif exige la normalisation et la standardisation d'une plate forme de travail [38] qui doit soutenir l'émergence et la coopération des organisations virtuelles [158] via l'exploitation des Technologies de l'Information et de la Communication (TIC). L'introduction des TIC vient considérablement améliorer les processus de décision dans les organisations pour introduire des systèmes coopératifs interactifs d'aide à la décision. [60] présente une approche d'intégration des nouveaux partenaires dans le réseau collaboratif. [45] présente des algorithmes opérationnels pour le choix et la sélection d'un partenaire.

Pour faciliter la coopération, les organisations ont besoin d'une infrastructure leur permettant de partager des documents, de travailler et de communiquer ensemble malgré les contraintes géographiques. C'est pourquoi les organisations virtuelles, les réseaux ou groupements d'entreprises s'appuient fortement sur les technologies de traitement de l'information.

Pour construire un système d'aide à la décision pour la gestion de la collaboration inter-organisations, les approches de recherche et d'extraction d'informations sont sollicitées pour découvrir l'information caractérisant le réseau [37] [131] [55]. Ces approches de recherche et d'extraction d'information ont vocation de devenir la pierre angulaire de systèmes d'information décisionnels, support de la gestion dynamique des cycles de vie de ces organisations collaboratives. D'un point de vue de recherche d'information on peut distinguer deux grandes catégories d'univers informationnels :

- Une recherche dans un environnement fermé où les organisations se mettent d'accord d'avance pour travailler ensemble à court terme (pour une durée précise). Pour ce faire, un certain nombre de partenaires prédéterminés partagent leurs connaissances et leurs informations (savoir faire, compétences...) sous un format donné et une structure homogène. Cette alliance est en général définie sur un court terme : une fois le bien ou le service livré, le partenariat prend fin. Ce type de réseau est caractérisé par des frontières très nettes et, dans le cycle de vie, les nouveaux venus ne sont autorisés qu'en cas d'incident (Exemple : un partenaire quitte le réseau).
- Une deuxième recherche qui se fait dans un environnement ouvert où les organisations ne se connaissent pas a priori et utilise une information hétérogène publique (par exemple l'information disponible sur le web). Ce type d'information rend la recherche plus difficile car on est face à des documents et des informations mal structurés. Par ailleurs, le processus collaboratif ciblé n'est pas connu non plus à l'avance, et le type de partenaires potentiels reste très ouvert. Notre travail se situe dans cette seconde approche, puisque nous allons nous intéresser à la recherche d'information au sein de sources publiques fournies par le web.

Dans le cycle de vie des organisations virtuelles on considère communément que la création réactive d'entreprises virtuelles à court terme requiert la mise en place préalable de réseau à long terme [172] classiquement dénommé VBE [39] [2]. Un « Virtual Organization Breeding Environment »(VBE) est défini comme une association d'organisations, adhérant à un accord de coopération à long terme et adoptant des principes et des infrastructures de fonctionnement communs. Un VBE est créé comme une association à long terme, de frontière nette, et ses membres sont recrutés dans un univers ouvert selon les critères définis par les créateurs ou des administrateurs. Une VO (Virtual Organization) est une organisation provisoire déclenchée par une occasion spécifique de collaboration. Ses associés sont principalement choisis parmi les membres du VBE. La création efficace de VOs dynamique exige un environnement approprié où les membres de nouveau VOs sont choisis selon des critères de possibilité et de confiance parmi elles. L'objectif principal du VBE est d'améliorer l'état de préparation de ses membres pour créer efficacement les VOs.

Au sein du cycle de vie des organisations virtuelles notre travail ne se situe pas sur la création d'entreprises virtuelles mais sur la création des VBE, qui intervient en amont. Dans cette optique les partenaires ne sont pas connus, il n'existe pas de relations de confiance préalables qui pourrait favoriser le partage d'informations privées. La création de VBE intervient par exemple dans l'analyse territoriale effectuée par des acteurs tels que les chambres de commerce ou autres. Dans ce contexte le nombre de partenaires potentiels à analyser est beaucoup plus large, ce qui rend pertinent une volonté d'usage de l'information publique et d'automatisation du processus à travers des mécanismes de recherche d'information bien spécifiques.

1.2.3 Deux contributions : Extraction d'informations sur les activités et sur les compétences

Au plan applicatif, nos recherches ont pour objectif de permettre l'intégration des mécanismes d'extraction d'information mis au point, au sein d'outils et de méthodes facilitant la construction de réseaux d'entreprises collaboratifs. Il est donc nécessaire de spécifier le besoin d'extraction d'information pour répondre à ce besoin final. Dans notre laboratoire, ce besoin final est étudié au sein de la communauté scientifique de Génie Industriel. Les informations susceptibles d'être utilisées en vue d'une aide à la décision pour la constitution de réseaux collaboratifs peuvent être variées. Il se pose donc la question du choix des informations clés, et du niveau de synthèse adéquat de ces informations.
Dans le cadre de nos recherches, nous avons fait le choix de nous référer aux travaux d'aide à la décision développés en Génie Industriel dans le laboratoire G2I, qui ont d'ores et déjà donner lieu à la publication de la thèse de M. Benali [15]. Outre la volonté d'approfondir les travaux précurseurs de Benali, le choix de cette approche d'aide à la décision est également motivé par 2 autres arguments : cette méthode est bien destinée à la constitution de réseaux à long terme de type VBE alors que la plupart des méthodes identifiées dans la littérature scientifique se positionnent

sur la création de VO (donc dans un contexte différent) ; par ailleurs, l'approche de Burlat et Benali [31] [15], présente le net avantage d'utiliser une information très synthétique sur les entreprises en vue de fournir une aide à la décision pertinente. Ainsi, la méthodologie d'analyse des réseaux d'entreprises proposée par [31] est basée sur une typologie des modes de coordination entre les différentes entreprises du réseau. Cette typologie est basée sur deux paramètres clés : la complémentarité des activités et la similarité des compétences. Ces deux paramètres ont été identifiés comme étant discriminants pour justifier le choix d'un mode de coordination industriel dans le cadre d'un groupement d'entreprises. Dans ses travaux, Benali n'a pas traité la question de l'accès aux informations sur la complémentarité des activités et la similarité des compétences. Au plan applicatif, notre thèse vise ainsi à proposer une automatisation des mécanismes d'extraction d'informations préalablement nécessaires à l'application des aides à la décision préconisées par Benali. En référence à cette méthode, nous allons donc nous intéresser à deux axes de recherches complémentaires : d'une part l'extraction d'information sur les domaines d'activités des entreprises et leur complémentarité ; d'autre part l'extraction d'information sur les compétences d'entreprises et leur similarité.
Sur le premier axe de recherche, l'objectif est d'arriver à détecter l'activité de l'entreprise à partir des données publiques, notamment son site web, pour établir un degré de complémentarité entre des secteurs d'activités distincts. Un premier système d'extraction d'information répondra à ce besoin. Sur le deuxième axe, l'objectif est d'arriver à établir une information synthétique correspondant à une similarité entre des ensembles de compétences caractérisant différentes entreprises. Cet objectif va induire le besoin d'extraire une trace des compétences d'une entreprise. Cette tâche d'extraction fait appel à la construction d'un deuxième système d'extraction d'information.

1.3 Démarche de recherche

1.3.1 Approche méthodologique adoptée

Notre travail est guidé par des objectifs et des hypothèses du génie industriel. La mise en place des solutions de traitement automatique de l'information est construite dans un but pragmatique décrit principalement par les objectifs d'aide à la décision et les types d'information nécessaires. Cette vision pragmatique crée un environnement spécifique où doivent être implémentés des solutions informatiques spécifiques pour atteindre les objectifs fixés. Nous avons commencé par une phase d'étude de la problématique générale et de l'état de l'art qui aboutit à identifier des besoins scientifiques et techniques associés concernant les activités et les compétences des entreprises. Ensuite nous avons appliqué une démarche systématique pour chacun de ces deux besoins de recherche et d'extraction d'informations. Cette démarche est composée en plusieurs étapes : analyse approfondie du contexte et des besoins de l'information, ciblage et développement d'une contribution conceptuelle, mise au point d'une application de test, préparation d'un corpus, analyse des

performances du système d'extraction sur ce corpus. Au final nous avons illustré comment les résultats des deux systèmes d'extraction d'information peuvent aboutir à l'application d'une procédure d'aide à la décision pour la construction de groupements collaboratifs d'entreprises.

Ainsi, l'ensemble de ce travail de recherche sera relaté dans le manuscrit en quatre parties successives. La première partie est consacrée à un état de l'art sur l'ensemble des méthodes et techniques de RI utiles à nos objectifs, en vue de dégager les contributions qui feront l'objet de la thèse. La seconde partie se focalise sur l'extraction d'information concernant les domaines d'activités des entreprises, en développant à la fois notre contribution conceptuelle et l'application sur un corpus afin d'analyser les performances du système. La troisième partie applique cette même démarche de recherche (soulignée ci-dessus), mais cette fois pour l'extraction d'information sur les compétences d'entreprises. Enfin la dernière partie de notre travail proposera une discussion des résultats et applications potentielles de ce travail, en dégageant des perspectives pour de futures recherches. La structuration détaillée de ces quatre parties est précisée dans la section suivante.

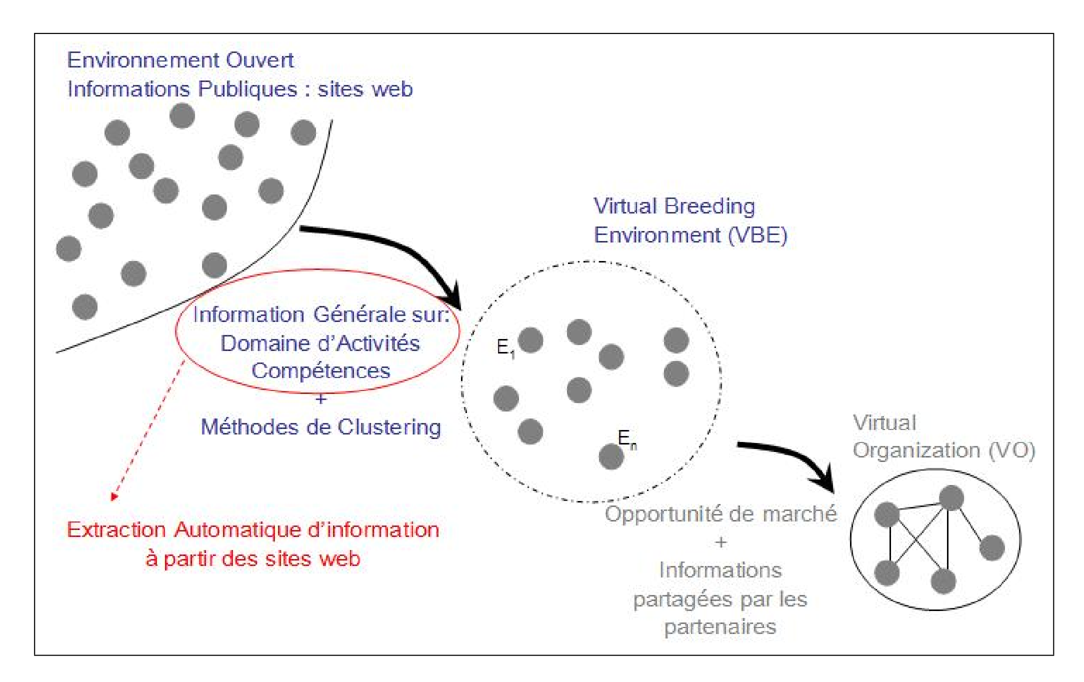


FIGURE 1.1 – La création des organisations virtuelles nécessite l'extraction d'information sur les domaines d'activités et les compétences

1.3.2 Organisation du mémoire

Ce mémoire est divisé en quatre parties :
La première partie fournit un état de l'art sur la littérature scientifique utile au ciblage et au développement de nos recherches. Elle comprend quatre chapitres :

le premier est une présentation générale du domaine de la recherche d'information qui définit le domaine de la recherche d'information documentaire, ses modèles et ses concepts de base. Ce chapitre justifie principalement l'approche proposée pour répondre à la question de la détection des activités des entreprises à partir de leur site web. Le deuxième chapitre définit et présente le domaine de l'extraction d'information, ses méthodes et ses principes. Le troisième chapitre est consacré à la définition des ontologies et à leur ingénierie, car ces notions seront utilisées dans la troisième partie de la thèse. Dans le quatrième chapitre, nous donnons un aperçu sur les techniques de traitement automatique de la langue (TAL). Nous distinguons dans un premier temps les niveaux d'analyse de la langue, avant de présenter le système UNITEX avec lequel nous réalisons le traitement linguistique.
Les deux parties qui suivent, présentent le cœur de nos contributions. La deuxième partie détaille notre proposition concernant la détection des activités d'entreprises à partir de leur site web : dans un premier chapitre, nous présentons la problématique des organisations virtuelles et le besoin de la recherche et l'extraction de l'information concernant les activités pour la construction des réseaux d'entreprises. Le second chapitre s'intéresse à la description de l'approche proposée pour la détection des domaines d'activités d'entreprises. Cette approche est basée sur l'indexation des sites web des entreprises en utilisant un vocabulaire hiérarchique contrôlé inspiré du NAF (Nomenclature des Activités Françaises) [3]. Cette contribution donne lieu à un premier système d'extraction d'information dont les performances sont analysées dans un troisième chapitre.

La troisième partie se situe dans la continuité de la précédente pour répondre à une question encore plus complexe d'extraction automatique d'informations sur les compétences d'entreprises. Elle comprend quatre chapitres. le premier explique le besoin d'extraction d'information sur les compétences d'entreprises lié au besoin d'une gestion efficace des compétences dans les réseaux d'entreprises. Un deuxième chapitre nommé "Ontologie des traces de compétences" décrit les différentes phases de construction et d'ingénierie d'une ontologie du domaine des compétences des entreprises. Cette ontologie est utilisée ultérieurement pour faire un traitement sémantique sur les textes des entreprises. Le troisième chapitre présente le processus de l'extraction d'information caractérisant les compétences des entreprises. Ce processus fait appel à l'ontologie du domaine et à un traitement linguistique basé sur les patrons syntaxiques qui décrivent les schémas structurels de l'information pertinente. Cette contribution d'extraction des informations sur les compétences a nécessité la création d'un logiciel, que nous avons baptisé UNICOMP. Dans un dernier chapitre une étude est faite sur la performance d'UNICOMP par rapport à l'information pertinente recherchée.

La quatrième partie est une synthèse des résultats des deux systèmes d'extraction. En s'appuyant sur les informations extraites par ces derniers, une maquette

3. http ://www.insee.fr/fr/nom-def-met/nomenclatures/naf/pages/naf.pdf

d'illustration sur la construction des réseaux d'entreprises est présentée. Dans un premier chapitre nous présentons l'application pour la construction des réseaux d'entreprises, notamment le calcul de la distance entre les différents ensembles des compétences relatives aux entreprises. Une cartographie, qui décrit les différents modes de coordination entre les entreprises mises en test, est présentée. Le deuxième chapitre nous permet de présenter une conclusion générale de nos contribution et d'évoquer les divers perspectives qui souvrent à la suite de notre travail de recherche dans le cadre de cette thèse.

Introduction Générale

Partie I : Positionnement et état de l'art

Chapitre II : Recherche d'Information
Chapitre III : Extraction d'Information et Fouille de Données
Chapitre IV : Les Ontologies
Chapitre V : Traitement Automatique de la Langue

Partie II : Détection Automatique des Activités d'Entreprises

Chapitre VI : Problématique
Chapitre VII : Détection Automatique des Activités d'Entreprises
Chapitre VIII : Application aux Réseaux d'Entreprises

Partie III : Extraction Automatique des Compétences d'Entreprises

Chapitre IX : Besoin d'Extraction
Chapitre X : Ontologie des Traces de Compétences
Chapitre XI : Extraction de Compétence
Chapitre XII : Performance du Système d'Extraction

Partie IV : Synthèse des résultats

Chapitre XIII : Application dans le Contexte des Réseaux d'Entreprises

Conclusion et Perspectives

Partie 1 : Positionnement et Etat de l'art

La recherche d'information (RI) est porteuse d'ambigüité : dans la vision Google, la recherche d'information est l'ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'utilisateur, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés. Cette recherche est connue sous le nom de la recherche documentaire. Un autre aspect de la recherche d'information fait appel à l'extraction d'information, la recherche d'entités nommées, les questions de classification, et donc le sens du domaine scientifique. Du concept de la RI se déclinent 3 sens :

1. Un sens intuitif dans la vie de tous les jours, qui vise à obtenir une information répondant à un besoin. Exemple « comment je déclare mes impôts sur internet ? »
2. Un sens plus scientifique, un peu plus précis mais qui reste très vaste et qui considère la recherche d'information comme un large domaine scientifique englobant différents buts (extraction d'information (EI), Classification, Question-Réponse (QR) etc.) :

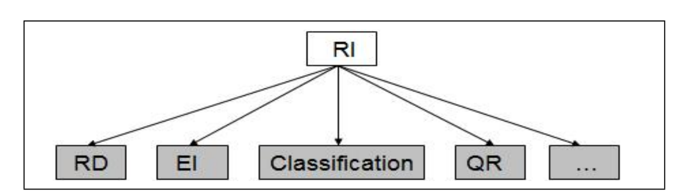


Figure 1.2 – Sens scientifique de la RI

3. Un sens plus restreint qui réduit la RI à la recherche documentaire (RD) telle qu'elle est développée dans le chapitre 2.

Dans cette partie, un état de l'art sera structuré en quatre chapitres sélectionnés par rapport à la problématique développée précédemment. Le premier chapitre présente la recherche d'information dans un sens réduit à la recherche documentaire, le deuxième chapitre décrit la recherche d'information dans un sens scientifique plus large. Dans le troisième et quatrième chapitre, nous détaillons respectivement les ontologies et le traitement automatique des langues, et comment ces deux domaines sont utilisées pour la recherche d'information.

Chapitre 2

Recherche d'Information

2.1 Introduction

> « *While a few centuries ago people were struggling to access information, today many are struggling to eliminate the irrelevant information that reaches them through various channels* »[Bukley, Berners Lee].

La croissance continue du volume de données (texte, image, vidéos) présentées sous différents formats, ainsi que l'apparition de disques offrant de gigantesques espaces de stockage ont imposé de définir des mécanismes pour gérer cette masse d'informations. Ce besoin a marqué la naissance du domaine de la « Recherche d'Information ». Depuis les années 1990, notamment avec l'avènement d'Internet, la recherche d'information est devenue un domaine important dans la communauté de la recherche scientifique. Aujourd'hui, la recherche d'information est un champ transdisciplinaire, qui peut être étudié par plusieurs disciplines, approche qui permet de trouver des solutions pour améliorer son efficacité. Elle met en jeu le stockage et la représentation de l'information d'une part, l'analyse et la satisfaction d'un besoin d'autre part.

Dans la littérature, les ouvrages consacrés à la recherche d'informations [139] [69] [10] la définissent comme l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents ceux qui sont susceptibles de répondre aux besoins de l'utilisateur. Ceci implique en général trois processus [13] : Poser une question (requête), construire une réponse (liste des documents pertinents), évaluer la réponse (jugements des documents restitués).

Le premier processus est lié au facteur cognitif humain. L'utilisateur a un besoin d'information pour acquérir une nouvelle connaissance absente chez lui ou compléter et confirmer une connaissance préalable. Parfois l'utilisateur est incapable de définir clairement son besoin d'information. A partir de cet état cognitif mal défini, l'utilisateur tente de s'exprimer dans le langage utilisable par le système, produisant ce que l'on nomme « Requête ». Le deuxième processus est la construction de la réponse par le système de recherche d'information. Ce dernier doit analyser la requête, prendre en compte les difficultés dues à l'ambigüité du langage naturel et présenter les solutions de façon compréhensible. Enfin le troisième processus, le jugement des documents pertinents, peut être défini comme une évaluation mentale par l'utilisateur de la réponse obtenue. En effet l'utilisateur peut être satisfait de la réponse retournée par le système de recherche d'information et dans ce cas le processus de recherche est arrêté, ou il est insatisfait et reformule sa requête pour lancer une nouvelle recherche.

2.2 Concepts de base de la RI

2.2.1 Le système de recherche d'information

Un système de recherche d'information (SRI) est l'interface entre la collection de données (corpus) et l'utilisateur qui attend une réponse pertinente après avoir lancé sa requête. Deux concepts principaux se déclinent autour de cette définition d'un Système de Recherche d'Information :

- Document : Les documents tels que manipulés par le SRI sont des documents logiques c'est-à-dire une unité du corpus, un ensemble d'information auto-explicative. Nous distinguons les documents logiques, et les documents physiques, ces derniers étant des fichiers sur un ordinateur. Il faut noter qu'un document logique peut être constitué de plusieurs fichiers informatiques, et réciproquement le corpus entier peut être contenu dans un seul fichier informatique. Par la suite nous utiliserons le terme « document »pour désigner un document logique.
- Requête : la requête est l'expression du besoin d'information de l'utilisateur. Elle est l'interface entre le SRI et l'utilisateur. Elle peut prendre plusieurs formes : ensemble de mots clés avec un ensemble des opérateurs (booléens par exemple), être exprimée en langue naturelle, etc.

Pour répondre au besoin d'information, le SRI fait appel à un ensemble de processus pour faire correspondre l'information contenue dans la collection de documents et le besoin d'information exprimé par la requête. Principalement, deux processus sont mis en œuvre :

- Processus d'indexation : c'est la transformation du document et de la requête en une représentation informatique qui reflète son contenu informationnel. Le résultat de l'indexation est un descripteur pour chaque document. Le plus souvent pour un document, ce descripteur contient une liste de termes auxquels sont associés des poids, qui tentent de caractériser le degré de représentativité de ces termes dans le document.
- Processus de recherche : c'est le processus noyau d'un SRI. Il permet d'associer à une requête l'ensemble des documents jugés pertinents par le système. Ce processus est lié au modèle de représentation de la requête et des documents. Il est basé sur un appariement entre la requête et les descripteurs des documents pour mesurer et évaluer leur pertinence. Les documents peuvent ensuite être classés selon cette évaluation de pertinence.

2.2.2 Indexation

L'indexation consiste à identifier l'information contenue dans tout le texte et à la représenter au moyen d'un ensemble d'entités, appelé index, pour faciliter la comparaison entre la représentation d'un document et d'une requête. Cette étape est primordiale pour la recherche dans des conditions acceptables de coût et d'efficacité. Dans la quasi-totalité des SRI, l'indexation est faite au niveau des termes ou des mots, que l'on appelle « mots clés »dans le sens où ils représentent l'essentiel de

l'information contenue dans les documents [146] [89] [49]. Cette indexation peut s'effectuer selon trois modes :

- Manuel : chaque document est analysé par un documentaliste ou un spécialiste du domaine et c'est lui qui attribue des mots clés aux documents.
- Semi-automatique : un processus automatique propose des mots clés et le choix final reste au spécialiste ou documentaliste.
- Automatique : un processus entièrement automatisé produit des mots clés.

L'avantage de l'indexation manuelle (ou semi-automatique) est qu'elle permet d'avoir une bonne correspondance entre les documents et les termes descripteurs. Ce qui améliore la précision dans les documents retournés par le système. En contrepartie, l'inconvénient de cette méthode est qu'elle exige un effort intellectuel en temps et en nombre de personnes, de plus le degré de subjectivité lié au facteur humain.
L'indexation automatique est celle qui a été la plus étudiée en recherche d'information. Il s'agit d'automatiser complètement la procédure d'indexation. On y distingue : l'extraction automatique des termes, l'utilisation d'un anti-dictionnaire pour éliminer les mots vides, la lemmatisation, le repérage des groupes de mots, la pondération des mots avant de créer l'index, etc. Le résultat de l'indexation est un ensemble de termes définissant ce qu'on appelle le langage d'indexation.

2.2.3 Pondération des termes

La pondération des termes est la détermination de l'importance des termes dans une requête ou un document. Autrement, c'est l'évaluation de leur pouvoir discriminant et leur importance dans la description sémantique du contenu d'un document. Pour déterminer ce pouvoir discriminant, on peut distinguer deux approches : la première est linguistique inspirée des techniques de traitement de la langue et la deuxième se base sur des aspects statistiques. Les techniques courantes de pondération de termes sont basées sur des notions de fréquence des termes dans un document *tf* (*term frequency* : mesure représentant l'importance locale d'un terme - fréquence relative) et de fréquence de ces termes dans l'ensemble des documents de la collection étudiée. On utilise l'inverse de cette mesure pour mesurer l'importance globale d'un terme *idf* (*inverse document frequency* - fréquence absolue). La mesure *idf* part du principe que « le nombre des documents pertinents à une requête est faible [en comparaison au nombre total des documents], et donc les termes apparaissant fréquemment doivent nécessairement apparaître dans beaucoup de documents non pertinents. En revanche, les termes peu fréquents ont une plus grande probabilité d'apparaître dans les documents pertinents et donc doivent être considérés d'une plus grande importance potentielle quand on cherche dans une base de données ». [1] [88].

1. « The number of documents relevant to a query is generally small, and thus any frequently occurring terms must necessarily occur in many irrelevant documents ; infrequently occurring query terms, conversely, have a greater probability of occurring in relevant documents and should thus be considered as being of greater potential importance when searching a database ».

2.2.4 Evaluation d'un SRI : Précision et Rappel

L'évaluation des systèmes de recherche d'information constitue une étape importante dans l'élaboration d'un modèle de recherche d'information. En effet, elle permet de caractériser le modèle et de fournir des éléments de comparaison entre modèles. La précision et le rappel sont deux éléments numériques qui permettent d'évaluer et de comparer des systèmes de recherche d'information. Appelons D l'ensemble des documents existants (le corpus), $Pert$ le sous-ensemble de D contenant tous les documents pertinents pour une requête q et $Retr$ le sous-ensemble de D contenant tous les documents retournés par le SRI, on a :

$$Precision = \frac{|Pert \cap Retr|}{|Retr|} Rappel = \frac{|Pert \cap Retr|}{|Pert|} \tag{2.1}$$

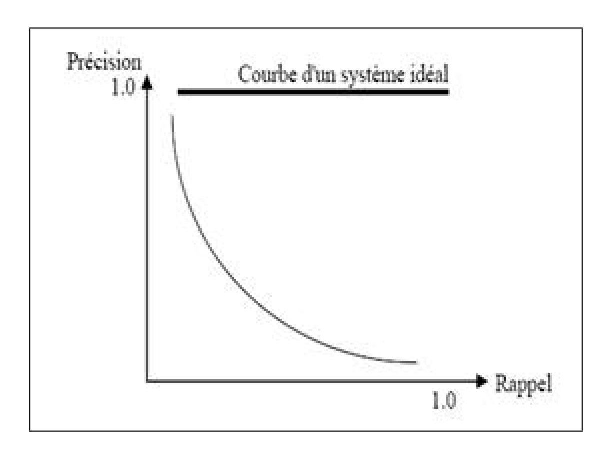


FIGURE 2.1 – Allure d'une courbe de Précision-Rappel

La précision est le pourcentage des documents retournés qui sont pertinents : c'est une estimation de la capacité du SRI à retourner des documents pertinents et donc à éliminer le bruit. Le rappel est le pourcentage des documents pertinents qui sont retournés par le SRI : c'est une estimation de la capacité du SRI à retourner tous les documents pertinents et à éliminer le silence.

$$Silence = Pert - Pert \cap Retr$$

$$Bruit = Retr - Pert \cap Retr$$

Le silence est l'ensemble des documents pertinents qui n'ont pas été retournés. Le bruit est l'ensemble des documents qui ne sont pas pertinents mais qui ont été retournés. Dans un système idéal, le taux de précision est égal à 1 à tous les niveaux de rappel. C'est-à-dire que tous les documents élus sont pertinents, et seuls ceux-ci ont été sélectionnés par le système. Dans ce cas on aura une droite. Ces indicateurs ne peuvent être mesurés qu'à partir d'un corpus parfaitement connu et maitrisé, c'est-à-dire que pour chaque requête, on connaît exactement les documents qui sont pertinents dans le corpus et qui doivent être inclus dans le résultat de la recherche. D'autre mesures d'évaluation existent telles que :

- La précision moyenne : elle prend en compte à la fois la précision et le rappel. C'est la moyenne des précisions calculées pour chaque document pertinent retrouvé, au rang de ce document. Si un document pertinent est retourné à la dixième position, la précision pour ce document est la précision à 10 documents.
- La R-précision : c'est la précision obtenue pour un nombre de documents retournés correspondant au nombre de documents pertinents dans la base.
- Le nombre total de documents pertinents retournés, ou le rappel à 1000 documents : ces mesures permettent d'évaluer la performance globale du système, en fonction ou non du nombre de documents pertinents total.
- Le rang du premier document pertinent : cette mesure a été proposée pour prendre en compte la satisfaction de l'utilisateur qui cherche un seul document pertinent.
- La longueur de la recherche *(Expected Search Length)* : elle est égale au nombre de documents non pertinents que doit lire l'utilisateur pour avoir un certain nombre n de documents pertinents.

La comparaison de deux systèmes de RI doit se faire sur le même corpus de test en utilisant la même mesure de performance. D'autres mesures telles que le temps de réponse ou la présentation des résultats peuvent être considérées, mais elles ne sont pas répandues à grande échelle, à cause de la difficulté de leur mise en œuvre. Les mesures basées sur Précision-Rappel restent les plus utilisées par les bancs d'essai les plus connus.

2.3 Les modèles de la RI

« A tentative description of a theory or system that accounts for all of its known properties »[Soukhanov, 84]

Un modèle de Recherche d'Information permet de fournir une formalisation du processus de recherche d'information. Il présente un cadre théorique pour la modélisation de la mesure de pertinence. Nous allons décrire ici les trois principaux modèles de la RI et particulièrement détailler comment se fait l'indexation des documents, comment se formulent les requêtes et comment s'effectue le calcul de la fonction de similitude.

2.3.1 Modèles booléens

Le modèle booléen est le modèle le plus simple, basé sur la théorie des ensembles et l'algèbre booléenne. Il propose une représentation de la requête sous forme d'une expression logique. Les termes d'indexations sont reliés par les connecteurs logiques $ET(\wedge)$, $OU(\vee)$ et $NON(\neg)$. Les poids sont naturellement binaires : si le terme existe dans le document alors son poids vaut 1, sinon il vaut 0. Une limite du modèle booléen est que pour une requête conjonctive donnée, il suffit qu'un seul terme ne soit pas présent dans un document pour que ce dernier soit considéré non-pertinent.

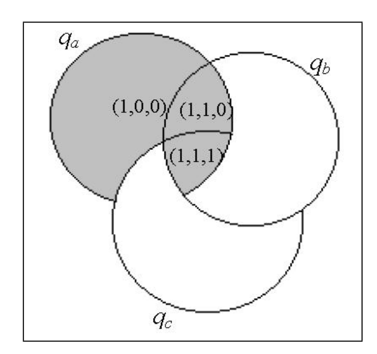


FIGURE 2.2 – Les 3 composants conjonctifs de la requête : $[q = q_a \wedge (q_b \vee \neg q_c)]$

De plus il n'y a aucun classement dans ce modèle : tous les documents considérés comme pertinents sont au même niveau de pertinence.
Pour remédier à ces limites, des extensions de ce modèle ont été proposés : le modèle booléen étendu [147] tient compte de l'importance des termes dans la représentation des documents et la requête en affectant des poids à chaque terme. [22] a proposé une extension du modèle booléen qui se base sur la théorie des ensembles flous. L'objectif de l'intégration des ensembles flous dans ce modèle est de réduire l'imperfection et de traiter l'imprécision qui caractérise le processus d'indexation, contrôler l'imprécision de l'utilisateur dans sa requête et traiter des réponses reflétant la pertinence partielle des documents par rapport aux requêtes. L'inconvénient majeur des modèles booléens est qu'ils ne sont pas adaptés au classement (ranking) des documents pertinents puisque les scores de pertinence sont calculés par des fonctions min et max qui ne couvrent pas nécessairement toutes les valeurs de pertinence des termes de la requête. Pour répondre à ce point, [22] [109] ont proposé des extensions qui prennent en compte l'aspect de l'ordonnancement des documents sélectionnés.

2.3.2 Modèles Vectoriels

Le modèle vectoriel a été développé par Salton [145] et ses collègues qui ont construit le système SMART (System for the Mechanical Analysis and Retrieval of Text) pour servir de base aux expériences d'IR. Une série de techniques d'IR (pondération de termes, classement, contrôle de pertinence) a été également conçue pour établir ce modèle. Indépendamment de la logique booléenne, le modèle de l'espace de vecteur a eu beaucoup d'influence sur le développement des systèmes opérationnels d'IR. Il présente une base unifiée pour d'éventuelles opérations de recherche, y compris l'indexation, le contrôle de pertinence et la classification de documents.
Dans le modèle vectoriel, chaque document D est représenté par un vecteur, avec N le nombre de ses descripteurs :

$$D_j = (d_1, d_2, d_3, ..., d_N),$$

Chaque requête est représentée par un vecteur :

$$Q = (q_1, q_2, q_3, ..., q_N),$$

Avec d_{ij} le poids du terme t_i dans le document D_j et q_i le poids du terme t_i dans la requête Q. Les poids sont des nombres positifs. Généralement ils représentent

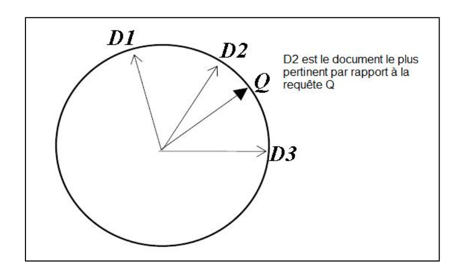


FIGURE 2.3 – Le modèle Vectoriel

l'importance du terme dans le document et dans l'ensemble des documents : si un terme se trouve souvent dans un même document, il représente bien ce document ; mais un terme qui apparaît dans tous les documents ne permet pas, à lui seul, de déterminer si le document est pertinent ou pas. Ces deux règles sont prises en compte dans le processus de pondération.
Soient N le nombre de documents dans un corpus et n_i le nombre de documents dans lesquels apparait le terme t_i. Soit tf_{ij} la fréquence du terme t_i dans le document d_i (plus t_i est présent dans d_j, plus tf_{ij} est grand). Soit idf_i la fréquence inverse du terme t_i dans l'ensemble des documents (D) (plus n_i est petit, plus idf_i est grand). On définit le poids d_{ij} par :

$$d_{ij} = tf_{ij} \times idf_i$$

Il y a différentes façons de calculer tf_{ij} et idf_i, nous présentons l'une des méthodes les plus courantes. Pour la fréquence on normalise le nombre d'occurrences du terme k_i dans un document d_j en le divisant par la valeur maximale d'apparition d'un terme dans ce document (on obtient un poids entre 0 et 1). Pour la fréquence inverse, on considère que si un mot est présent dans tous les documents, sa fréquence inverse est nulle, sinon elle est calculée comme le logarithme du quotient de N et n_i.

$$tf_{i,j} = \frac{freq_i, j}{max(freq_{i,j})}$$

$$idf_i = log\frac{N}{n_i}$$

Pour évaluer le degré de similitude entre le document et la requête, on calcule la corrélation entre les deux vecteurs. Les principales mesures de similarité sont :
Le produit scalaire :

$$RSV(Q, D_j) = \sum_{i=1}^{N} q_i \times d_{ij}$$

La mesure de Jaccard :

$$RSV(Q, D_j) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{\sum_{i=1}^{N} q_i^2 + \sum_{i=1}^{N} d_{ij}^2 - \sum_{i=1}^{N} q_i \times d_{ij}}$$

La mesure cosinus :

$$RSV(Q, D_j) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{(\sum_{i=1}^{N} q_i^2)^{(1/2)} \times (\sum_{i=1}^{N} d_{ij}^2)^{(1/2)}}$$

Ce modèle est utilisé aussi pour le QBE (*Query By Example, ou recherche par l'exemple*) [173], dont le principe est de fournir au système un document pertinent (un exemple) pour qu'il recherche les autres documents pertinents.

2.3.3 Modèle Connexionniste

Ce modèle se base sur le formalisme des réseaux de neurones [97], [98], [26] [119]. Un réseau est établi à partir des représentations initiales des documents et de l'information descriptive associée (mots clés). Cette approche permet de passer d'une simple comparaison des requêtes et des documents aux techniques basées sur des associations sémantiques entre les termes pour l'expansion de la réponse. Le processus d'appariement est basé sur la propagation de signaux entre la couche d'entrée et la couche de sortie (voir la figure 1.4). Chaque neurone de la couche d'entrée calcule une valeur et la transmet aux neurones de la couche suivante. Ce processus se reproduit jusqu'à l'arrivée à la couche de sortie. Les réseaux de neurones sont

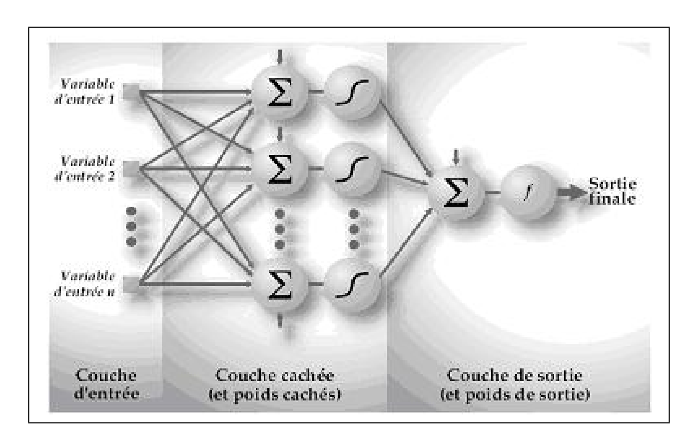


FIGURE 2.4 – Typologie d'un réseau de neurones

composés d'éléments simples appelés neurones, fonctionnant en parallèle. Un neurone est un processeur qui applique une opération simple à ses entrées et que l'on peut relier à d'autres pour former un réseau qui peut réaliser une relation entrée-sortie quelconque. Ces éléments ont été inspirés par le système nerveux biologique. Le fonctionnement du réseau de neurone est fortement influencé par la connexion

des éléments entre eux. On peut entraîner un réseau de neurone pour une tâche spécifique (classification par exemple) en ajustant les valeurs des connexions (ou poids) entre les neurones.
Le modèle de [120] est l'un des premiers systèmes qui développe le connexionnisme pour la RI, il comprend deux types de cellules, les cellules termes et les cellules documents, en utilisant les liens inhibiteurs pour réaliser un réseau de type *Winner take all* qui ne permet de retourner qu'un seul document. En 1991, Lin et ses collègues [108] ont présenté leur modèle dans lequel, pour chaque document, on associe un vecteur où les coordonnées correspondent aux termes représentatifs. Ce vecteur est l'entrée du réseau qui va chercher le neurone gagnant (neurone le plus actif) et renforce son poids ainsi que les neurones les plus proches. Une des limites de ce modèle est que la pertinence d'un document est fortement liée à la représentation initiale de ses termes représentatifs. Dans un modèle plus évolutif, des techniques de retour de la classification de l'utilisateur et de la mémorisation de la connaissance pour la reformulation automatiques des requêtes ont été utilisées. Les cellules représentent deux types d'information : les documents et les termes du langage d'interrogation. Les connexions sont basées sur des liens d'association sémantique entre les termes, des associations de synonymie, c'est-à-dire qu'un terme peut remplacer l'autre, les associations de généricité/spécificité permettent d'exprimer quelque chose de plus général ou de plus précis suivant les concepts des termes, ainsi que sur des liens de co-occurrence qui regroupent les termes qui apparaissent ensemble lors de l'indexation d'un document. Ces associations ont des sens différents. Pour cette raison, elles sont représentées par des liens de pondérations différents et elles peuvent être combinées ou bien utilisées séparément.
Le modèle de Boughanem [26] vise essentiellement à résoudre les problèmes posés par les approches classiques. Il propose une méthode pour tenir compte des relations qui peuvent exister entre les termes. Ce modèle utilise une représentation connexionniste et dynamique avec un réseau de deux couches. Des techniques d'apprentissage ont été mises en place pour améliorer les performances. D'autre part l'expansion ou la reformulation des requêtes intègrent de nouveaux termes.
Dans ce modèle, un critère clé pour fonder les relations inter-termes est l'occurrence dans la base des documents. Plusieurs fonctionnalités ont été mises en place pour améliorer la pertinence, avec un système dynamique qui évolue suivant les exigences de l'utilisateur ;

- Trouver la possibilité de reformuler la requête suivant la connaissance de la base des documents.
- Tenir compte des requêtes précédentes pour la réorganisation de la base.

2.3.4 Modèle Probabiliste

La recherche d'information a été également influencée par la théorie mathématique des probabilités pour définir un modèle probabiliste [96] [142] [141]. La pertinence d'un document par rapport à une requête correspond à un degré de probabilité de pertinence. Pour ce faire, le processus de décision complète le procédé d'indexa-

tion probabiliste en utilisant les deux probabilités conditionnelles suivantes :
$P(t_i/Pert)$: probabilité que le terme t_i apparaisse dans un document donné sachant que ce document est pertinent pour la requête.
$P(t_i/NonPert)$: probabilité que le terme t_i apparaisse dans un document donné sachant que ce document n'est pas pertinent pour la requête.
En utilisant la formule établie par Bayes et en supposant l'indépendance des variables "document pertinent" et "document non pertinent", la fonction de recherche peut être obtenue en calculant la probabilité de pertinence $P(Pert/D)$ d'un document D donné [142] [139].
Soit $D(t_1, t_2, ..., t_N)$ où
$t_i = 1$ si le terme t_i indexe le document D, sinon $t_i = 0$
$P(Pert/D) = \frac{P(D/Pert).P(Pert)}{P(D)}$ et $P(NonPert/D) = \frac{P(D/NonPert).P(NonPert)}{P(D)}$ Avec :
$P(Pert/D)$ est la probabilité de pertinence d'un document sachant sa description.
$P(D) = P(D/Pert).P(Pert) + P(D/NonPert).P(NonPert)$
$P(D/Pert)(respectivement P(D/NonPert)))$ est la probabilité d'observer D sachant qu'il est pertinent (respectivement non pertinent).
$P(Pert)(respectivement P(NonPert))$ est la probabilité a priori pour qu'un document soit pertinent (respectivement non pertinent).
Pour la restitution, les documents sont rangés en fonction de $P(Pert/D)$. Le principe d'ordonnancement probabiliste entraine que cet ordonnancement est optimal en ce sens que, quelque soit le pourcentage de documents qui sont restitués, le pourcentage de documents restitués qui sont pertinents est maximisé. les systèmes Okapi [140] et Inquery [34] reposent sur ce modèle.

2.4 Conclusion

Dans ce chapitre, nous avons présenté le domaine de la recherche d'information (recherche documentaire) et ses techniques. Ces dernières seront sollicitées dans notre travail pour résoudre la première problématique de détection automatique du secteur d'activités d'une entreprise à partir de son site web.
Dans la deuxième problématique qui est l'identification des compétences des entreprises, nous avons besoin d'explorer et de comprendre les notions du domaine de l'extraction d'information parce que la réponse recherchée n'est pas un document. C'est plutôt une information précise d'un domaine spécifique. C'est pourquoi la partie suivante de ce mémoire s'oriente vers cet objectif.

Chapitre 3

Extraction d'Information et Fouille de Données

3.1 Introduction

La réponse à la détection des compétences des entreprises n'est pas de retourner un document ou un texte approximatif. Il s'agit de relier les éléments pour construire l'information complète et structurée à partir d'une fouille et d'une compréhension des données. L'objectif de ce chapitre est de définir l'extraction d'information et les méthodes de fouille de données, à savoir la fouille de texte, en présentant à chaque fois des exemples de systèmes qui ont été implémentés pour donner une représentation sémantique profonde du texte.

3.2 Extraction d'information

3.2.1 Définition

- *« Extraction information is the name given to any process which selectively structures and combines data which is found, explicity stated or implied, in texts »* [Wills, 97]
- *« L'extraction d'information désigne l'activité qui consiste à remplir automatiquement une banque de données ou encore un formulaire à partir de textes en langage naturel »*[Pazienza, 97]
- *« L'extraction d'information est l'activité qui consiste à remplir une source de données structurées (base de données) à partir d'une source de données non structurées (texte libre) »* [Gaizauskas et al., 98]

L'extraction d'information s'oppose classiquement à la recherche d'information qui vise à retrouver, dans une collection de documents, un sous-ensemble de documents pertinents vis à vis d'une requête. L'extraction d'information nécessite une analyse du texte pour interpréter et construire une représentation formelle. C'est une tâche difficile qui requiert une part de compréhension et nécessite des connaissances, des ressources lexicales, sémantiques et conceptuelles adaptées aux documents et au domaine à traiter pour restituer une information complète et structurée.

3.2.2 Systèmes d'extraction d'information

Les systèmes d'extraction actuels ont bénéficié de l'apport des systèmes de compréhension traditionnels. La compréhension de texte est un domaine exploré depuis

le début du traitement des langues [144] ; dans les années 1960, on a assisté à la création de modèles visant à rendre compte du contenu des documents pour la recherche documentaire (système KWIC, recherche statistique des mots les plus significatifs) [147]. Durant les années 1970, des systèmes plus perfectionnés pour l'interrogation en langage naturel de base de données sont apparus, comme le système Lunar grâce auquel, au retour des missions Apollo, les géologues pouvaient interroger en anglais la base des minéraux collectés sur la lune. La compréhension de texte est définie comme :

- L'extraction de toute l'information du texte, qu'elle soit pertinente ou non.
- La compréhension du discours et les nuances de sens.

Quelle que soit l'architecture des systèmes de compréhension de texte, l'objectif est toujours le même : donner au texte une représentation sémantique profonde.

3.2.2.1 Le système Kalipos

C'est un système question/réponse en langue naturelle qui a été développé au centre scientifique d'IBM France à partir de 1985 [30]. Ce système permet de produire des graphes conceptuels : l'analyse syntaxique est réalisée à l'aide d'une grammaire contextuelle implémentée sur un modèle proche des grammaires à clauses définies (DCG). Cette représentation est relativement proche de la surface : par exemple elle n'effectue qu'un traitement limité du passif. Au moment de sa construction des graphes conceptuels, c'est la partie sémantique de l'analyseur qui se charge de ces questions. Au début des années 1990, Kalipos a été intégré dans deux projets, Menelas et Exosème.

3.2.2.2 Le système Menelas

Dans ce système [174], la première partie de l'analyse de chaque phrase est effectuée par Kalipos. La partie sémantique de Kalipsos est réduite à la production d'un graphe dont le contenu est proche de la structure syntaxique profonde de la phrase analysée. La composition sémantique est effectuée dans un second temps, dynamiquement, en s'appuyant sur un modèle élaboré des connaissances du domaine.

3.2.2.3 Le projet Lilog

C'est un projet d'IBM-Allemagne pour le développement d'un système de compréhension de textes écrits en allemand. Les traitements sémantiques de ce système sont réalisés en trois étapes : une première analyse compositionnelle, à partir du formalisme syntaxique HPSG, qui traite en particulier des relations actancielles ; un traitement complémentaire, basé sur des représentations sémantiques inspirées de la DRT *Discourse Representation Theory*, qui traite entre autres de la temporalité et des anaphores ; enfin un module de raisonnement utilisant une logique typée qui intègre les connaissances sur le domaine. Ce système a été mis au point et testé pour une application de tourisme dans la ville de Düsseldorf.

Ces systèmes de compréhension de texte présentés ci-dessus s'appuient sur des théories et formalismes syntaxiques (grammaires transformationnelles, grammaire d'unification, etc.) et sémantiques (DRT, graphes conceptuels, etc.) qui ont été étudiés de manière approfondie, les aspects pragmatiques et contextuels étant moins développés et formalisés. L'inconvénient de ces systèmes est que leur adaptation pose des problèmes de réutilisation. Pour l'application à une nouvelle tâche, ils nécessitent la reconstruction d'une grande partie de la base de connaissances et du lexique sémantique. Cette opération est dans la majorité des cas manuelle, coûteuse et peu reproductible, car toutes les règles et heuristiques de l'analyseur doivent être mises à jour puisqu'il ne tient compte que d'un sous langage donné.

3.2.3 Evaluation des systèmes d'extraction d'information

L'évaluation des systèmes d'extraction d'information consiste à déterminer le bruit (information extraite de manière erronée) et le silence (information pertinente non extraite). C'est ce type d'évaluation que proposent les conférences américaines *Message Understanding Conférences* (MUC).
MUC est une conférence internationale d'évaluation de systèmes de compréhension automatique de messages en langue naturelle. Elle est organisée par le Département de la Défense, l'ARPA (Advanced Research Projects Agency, États-Unis). Les participants à cette campagne doivent développer un système capable d'extraire le maximum d'informations pertinentes d'un corpus d'entrainement. Ce corpus est diffusé à l'avance avec la liste des informations à identifier :

- MUC 1 (1987) et MUC 2 (1989) ont analysé les rapports d'opérations tactiques navales.
- MUC 3 (1991) et MUC 4 (1992) avaient pour objectif d'analyser des textes journalistiques traitant du terrorisme en Amérique Latine, afin d'extraire des dépêches de presse la maximum d'informations sur des actes terroristes comme le nom des groupes, le nom des victimes, les types d'armes, les dates et les lieux, etc.
- MUC 5 (1993) traitait un corpus économique : fusion, rachat et création d'entreprises internationales de fabrication de circuits électroniques.
- MUC 6 (1995), constituait une suite de MUC 5 : traitement des changements de dirigeants à la tête des entreprises.
- MUC 7 (1998), analyse de textes journalistes rapportant des crashs d'avion et de tirs de missiles.

Le département américain de défense a favorisé l'émergence de nouvelles compagnes d'évaluation, pour poursuivre l'évolution des différents systèmes de traitement de textes :

- TIPSTER : c'est un programme pour l'évaluation des systèmes de résumé de textes. Un lien a été établi entre MUC et TIPSTER pour définir une architecture standard de traitement de textes écrits. Cette action a débouché sur la définition de formats d'annotations et d'interfaces de programmation (API)

standards.

– TREC (*Text Retrieval Conference*) : c'est une conférence qui rassemble les concepteurs de boîte à outils et de logiciels de recherche d'information sur des documents plats. Elle permet la comparaison des performances des systèmes, sur des volumes importants de données (plus de 500 Mo de données dès TREC 1). Ces conférences sont vues comme un standard international, annuel, dans le domaine de l'évaluation de la recherche d'information.

Pour mieux comprendre ce domaine (EI), nous allons expliciter le plus largement possible le domaine de la fouille de données, ses méthodes et ses dérivations qui seront nos bases pour répondre à nos objectifs d'extraction.

3.3 Fouille de données

C'est un processus non trivial d'extraction et d'analyse de lots importants de données dans le but de décrire des tendances passées, de prédire des tendances futures et/ou d'extraire une information pertinente. Historiquement, le terme « fouille de données »ou « data mining »a été employé en statistique et en bases de données pour désigner des données impropres, non prêtes à l'analyse et qui nécessitent une phase de nettoyage. A l'origine la fouille de données fait partie d'un domaine plus large appelé extraction de connaissances dans des bases de données (ECBD) ou *Knowledge discovery in databases* (KDD) [46] [73].

3.3.1 Extraction de connaissances dans des données (ECD)

« L'ECD désigne le processus non trivial conduisant à la découverte des informations implicites, inconnues jusqu'alors et potentiellement utiles et compréhensibles à partir de données » [Piatesky-Shapiro et al., 91]

« KDD is the nontrivial process of identifying valid novel, potentially useful, and ultimately understandable patterns in data » [Fayyad, 96]

« Because computers have enabled humans to gather more data than we can digest, it is only natural to turn to computational techniques to help us unearth meaningful patterns and structures from the massive volumes of data. Hence, KDD is an attempt to address a problem that the digital information era made a fact of life for us all : data overload. » [Fayyad, 96]

L'extraction des connaissances à partir des données ECD est un processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par un utilisateur qui y joue un rôle central [58]. L'interactivité est liée aux différents choix que l'utilisateur est amené à effectuer, car l'ECD est composée de plusieurs tâches et l'utilisateur peut décider de revenir en arrière à tout moment si les résultats ne lui conviennent pas.

Le pré-traitement : consiste à sélectionner et transformer les données utiles à une

problématique de manière à les rendre exploitables par un outil de fouille de données.
La fouille de données : c'est le cœur et l'étape la plus complexe dans le processus d'ECD, elle consiste à appliquer des méthodes intelligentes dans le but d'extraire des motifs. Ces motifs correspondent à l'expression, dans un langage donné, d'un sous-ensemble de données recherchées.
L'évaluation et la présentation : consiste à mesurer l'intérêt des items générés et à les présenter à l'utilisateur grâce à différentes techniques de visualisation.

3.3.2 De la fouille de données à la fouille de texte

« Text mining is the science of extracting information from hidden patterns in large textual collections » [Feldman, 98]

« I'd like suggesting defining KDT rather as the science that discovers Knowledge in texts, where "knowledge" is taken with the meaning used in KDD, that is : the knowledge extracted has to be grounded in real world, and will modify the behaviour of a human or mechanical agent. » [Kodratoff, 99]

Alors que la « fouille de données »traite des bases de données structurées, la « fouille de texte »ou *text mining* traite des textes. C'est la réponse actuelle au problème de la surcharge informationnelle de type textuel (il est admis que les textes constituent l'essentiel de l'information (80%) disponible dans les mémoires électroniques).
La fouille de texte permet de découvrir des connaissances éventuellement cachées dans de très volumineuses données textuelles. Ce travail consiste à extraire des corrélations entre les différentes entités. Pour ce faire, deux approches existent : une approche basée sur le traitement automatique des langues, propre à la nature des données à traiter et une approche statistique (analyse des données) pour corréler les données entre elles, et en saisir les invariants et les règles qui les régissent.

3.3.3 Système de fouille de texte

Des systèmes de résumé automatique basé sur des techniques de fouille de texte ont vu le jour ; le système CAST *Computer Aided Summarization Tool* [33] est un système de résumé automatique reposant sur une approche semi-automatique qui prend en entrée un texte étiqueté. Il intègre plusieurs méthodes de séléction de phrases importantes, dont la mesure $tf.idf$, des indices positionnels (position et longueur des phrases) des indices récursifs et la cohésion lexicale. Le système Sygmart développé par Yousfi-Monod et Prince [170] repose sur une analyse syntaxique des phrases. L'idée de son approche est d'épurer les phrases d'un texte de ses compléments circonstanciels de lieu, de temps ou de manière pour obtenir un texte comprimé réduit à ses éléments essentiels qui sera le résumé. D'autres systèmes dédiés pour la détection des tendances émergentes, qui est une tâche importante dans la veille scientifiques et technologiques, ont été développés. Le système HDDI *Hirarchical Distributed Dynamic Indexing*[95] a pour objectif de regrouper les documents

dans des régions de « sous-thématiques de similarité sémantique »pour générer automatiquement une hiérarchie de sujet afin d'organiser les documents à la manière des taxonomies des moteurs de recherche ou des annuaires de type Yahoo. TermWatch [149] est un système de classification automatique qui vise à cartographier les thèmes d'un corpus. Son originalité réside dans le fait que pour identifier les sujets majeurs dans un ensemble de textes, les unités textuelles peuvent être agrégées selon d'autres dimenssions que la co-occurrence. Il est capable de regrouper les termes en fonction des relations de variation internes sans prendre en compte leurs co-occurrences dans les documents.

3.3.4 Quelques méthodes de fouille de données

L'objectif des méthodes de fouille des données est de rechercher des similarités ou des relations de dépendance entre les ensembles des unités qui constituent le corpus. Elles sont issues du croisement entre la statistique (analyse des données), d'intelligence artificielle (méthodes d'apprentissage) et de bases de données. Les méthodes d'analyse de données se répartissent en deux grandes familles selon la tâche à effectuer [85] : les méthodes descriptives et les méthodes prédictives. Ces méthodes d'analyses représentent des spécialités de recherche à part entière, dont il serait hors de propos de faire une présentation exhaustive dans ce mémoire ; c'est pourquoi nous nous arrêtons sur leurs principes fondamentaux.

3.3.4.1 Méthodes descriptives

Les méthodes descriptives ont pour but de proposer une structure à partir d'un ensemble de données, en l'absence d'une structure cible existante. Elles sont non supervisées parce qu'elles n'ont pas au départ un modèle des données ou un modèle de la structure cible à trouver. Elles sont beaucoup utilisées pour les tâches de classification automatique pour faire émerger la structure sous-jacente à un ensemble de données [91].

Méthodes de classification automatique Les méthodes de la classification automatique se déclinent en deux types :

Méthodes hiérarchiques

Historiquement, ce sont les premières développées, en raison de la simplicité des calculs. L'avènement de puissants ordinateurs leur a fait perdre une certaine popularité au profit des méthodes non-hiérarchiques. Toutefois dans certains domaines (comme la paléontologie), elles demeurent d'utilisation courante en raison de leur capacité d'organiser des ressemblances suivant une hiérarchie. Elles consistent à former automatiquement des classes d'objets. La construction de la hiérarchie peut être ascendante (Classification Ascendante Hiérarchique CAH) ou descendante (Classification Descendante Hiérarchique CDH). L'algorithme de base pour une CAH classification est le suivant :

- Tant qu'on a plus d'un groupe,
- Calculer les ressemblances entre toutes les paires de groupes,

– Fusionner les deux groupes montrant la plus grande ressemblance (similarité) ou la plus faible dissemblance (dissimilarité).

Les méthodes hiérarchiques diffèrent entre elles par le choix du critère de ressemblance et la façon de mesurer les ressemblances entre un nouveau groupe fusionné et les autres inchangés.

Méthodes de partitionnement

Elles correspondent à une famille d'algorithmes de classification connus sous le nom générique de *k-means* [110]. Contrairement aux méthodes hiérarchiques largement basées sur les mesures de similarité, les méthodes de k-means nécessitent des mesures de distance pour déterminer la distance qui sépare les individus à classer. Un algorithme de k-means procède de la manière suivante :

1. l'utilisateur choisit le nombre k de classes à former.
2. les objets à classer (unités textuelles, documents) sont répartis aléatoirement dans des classes par l'algorithme.
3. l'algorithme calcule le centroïde de chaque classe. Le centroïde de la classe est son point d'équilibre, qui se trouve à équidistance de tous les autres points de la classe et qui sera le représentant de la classe.
4. la distance qui sépare chaque individu du centroïde d'une classe est calculée et un individu est affecté au centroïde dont il est le plus proche.
5. l'algorithme recalcule le nouveau centroïde de chaque classe. Tant que les individus changent de classe, ou tant que les centroïdes changent, les étapes 3 à 5 sont réitérées, sinon l'algorithme s'arrête.

Les méthodes factorielles C'est une série de méthodes d'analyse des données dont l'objectif est de représenter sur un plan 2D les proximités/distances observées entre les lignes et les colonnes dans un tableau de contingence. Les deux méthodes les plus utilisés en analyse des données textuelles sont l'*Analyse Factorielle des Correspondances* (AFC) et le *Latent Semantic Analysis* (LSA).

Analyse factorielle des correspondances :

L'analyse factorielle des correspondances a été développée par Benzécri en 1973 ; sa problématique est la suivante : comment reproduire les distances observées entre les points lignes et les points colonnes d'un tableau de contingence sur un espace 2D tout en diminuant la perte ou la déformation d'information entre elles ? Pour mesurer la distance entre chaque point ligne et chaque point colonne, il a utilisé la distance du $chi - 2$ [114].

L'hypothèse de cette distance est que si deux lignes i et i' ont la même distribution c'est-à-dire le même profil et qu'on remplace les deux lignes i et i' par une nouvelle ligne i'', somme des deux précédentes, la distance entre la nouvelle ligne i'' et deux colonnes j et j' ne doit pas être modifiée. La distance du $chi - 2$ s'écrit :

$$d^2(i,i') = \sum_j \left(\frac{1}{f.j}\right)\left(\frac{f_{ij}}{f.j} - \frac{f_{i'j}}{f'_i}\right)^2$$

Cette formule stipule que le carré de la distance d entre deux points lignes i et i' ou entre deux points colonnes j et j' est égale à la fréquence relative du point ligne

i dans la colonne j moins la fréquence relative du point i' dans la colonne j.
Analyse sémantique latente (Latent Semantic Analysis, LSA) :
Le principe général de l'analyse sémantique latente [100] est de définir la signification des mots à partir des contextes dans lesquels ils apparaissent au sein de vastes corpus de textes. Le LSA prend en entrée une matrice croisant en ligne les objets d'étude (les mots), et en colonnes les contextes dans lesquels ils apparaissent (le texte, le paragraphe ou la phrase). Chaque case contient le nombre d'occurrences d'un mot dans un contexte. Le LSA trouve des applications dans des domaines très variés, sa variante initiale est le LSI (*Latent Semantic Indexing*) qui est appliquée à la problématique de la recherche d'information.
Les règles d'associations La méthode des règles d'association [3] est une méthode d'apprentissage non supervisé (apprentissage qui se base sur des lois locales et ne nécessite pas une intervention ou une règle de l'utilisateur, on laisse le système s'auto-organiser). Elle permet de découvrir, à partir d'un ensemble de transactions, un ensemble de règles qui expriment une possibilité d'association entre différents items (mots, attributs, concepts). Une transaction est une succession d'items exprimés selon un ordre donné ; de même l'ensemble des transactions contient des transactions de longueurs différentes.
Une règle d'association est une implication de la forme $X \Rightarrow Y$ où X, Y appartient à I avec $I = i_1, i_2, i_3...i_n$ un ensemble d'items.
Pour une règle d'association $X \Rightarrow Y$ on définit le support S et la confiance C.
Le support d'une règle $X \Rightarrow Y$ est : support$(X \Rightarrow Y)$ = support$(X \cup Y)$.
La confiance d'une règle est le rapport entre le nombre de textes contenant $X \cup Y$ et le nombre de textes contenant X, ce qui reflète la probabilité conditionnelle $P(X/Y)$. Lorsque la confiance vaut 1 la règle est dite exacte, sinon elle est approximative.
La confiance d'une règle $X \Rightarrow Y$ est : confiance $(X \Rightarrow Y)$ = support$(X \cup Y)$ /support(X). Notons que confiance $(X \Rightarrow Y) \in [0; 1]$. C'est la proportion des objets qui possèdent à la fois X et Y parmi les objets qui possèdent déjà X.
Le support et la confiance sont des mesures d'intérêt définis par l'utilisateur. Ces deux mesures permettent de réduire le nombre des règles extraites. Le support d'une règle est donné par le nombre de textes contenant à la fois les termes clés de X et Y.
[12] ont étudié les questions soulevées par l'utilisation de ces deux mesures pour indiquer la force d'une règle. Le tableau suivant 3.1 montre les différents cas de la combinaison de ces deux paramètres.

Les règles d'association jouent un rôle très important dans la découverte de la compréhension des relations de dépendances entre les variables dans une base de données. Elles sont utilisées aussi pour la prédiction face à une nouvelle instance : le conséquent est la cible de la prédiction (l'événement qui doit se produire). Le degré de confiance d'une règle indique également le degré de confiance de la prédiction.

	Taux de confiance bas	Taux de confiance élevé
Support élevé	La règle est rarement juste mais peut être utilisée fréquemment	La règle est souvent juste et peut être utilisée fréquemment
Support faible	La règle est rarement juste et ne peut être utilisée que rarement	La règle est souvent juste mais ne peut être utilisée que rarement

Table 3.1 – Compromis possible entre Support et Confiance d'une règle [12]

3.3.4.2 Méthodes prédictives

Ces types de méthodes d'analyse de données permettent de prédire la catégorie d'un futur objet à classer, par le biais d'une phase d'apprentissage. Elles exploitent les résultats issus des recherches en apprentissage machine, en probabilités et en sciences cognitives.
k-plus proches voisins Connue en anglais sous le nom *k-nearest neighbor* (K-NN) [44]. Cette méthode diffère des méthodes traditionnelles d'apprentissage car aucun modèle n'est induit à partir des exemples.
Pour prédire la classe d'un nouvel élément, l'algorithme cherche les k plus proches voisins de cet élément. La méthode utilise donc deux paramètres : le nombre k et la fonction de similarité pour comparer le nouvel élément aux éléments déjà classés.
Tout d'abord, l'algorithme doit produire, pour chaque catégorie, un modèle qui associe des poids pour chaque document et une valeur seuil pour décider de l'appartenance d'un document dans une classe. [168] a étudié différentes stratégies pour déterminer ce seuil. Ensuite il faut choisir la taille des k voisins qui peut varier en fonction des données dans le contexte de l'objet à classer. La taille de k est déterminée empiriquement par plusieurs essais. Enfin, il faut choisi parmi les mesures de similarité existantes celle qui sera utilisée pour comparer un nouvel objet aux cas déjà classés.
Les machines à vecteurs supports Le principe des machines à vecteurs supports (*Support Vector Machine SVM*) [159] (ou *séparateur à vaste marge*) suppose que l'on peut séparer linéairement les classes dans l'espace de représentation des objets à classer. En d'autres termes l'objectif est de trouver une surface linéaire de séparation (hyperplan) maximisant la marge entre les exemples positifs et négatifs d'un corpus d'apprentissage. La distance séparant les vecteurs les plus proches de l'hyperplan doit être maximale. Ces vecteurs sont appelés « vecteurs supports ». Un nouvel objet est classé en fonction de sa position par rapport à l'hyperplan. La méthode SVM est plus coûteuse en temps d'apprentissage [87] que les classifieurs bayésiens naïfs ou k-plus proches voisins, cependant elle donne de bons résultats pour la classification de textes [106].
Les classifieurs bayésiens naïfs Ces classifieurs se fondent sur le théorème de Bayes énoncé comme suit :

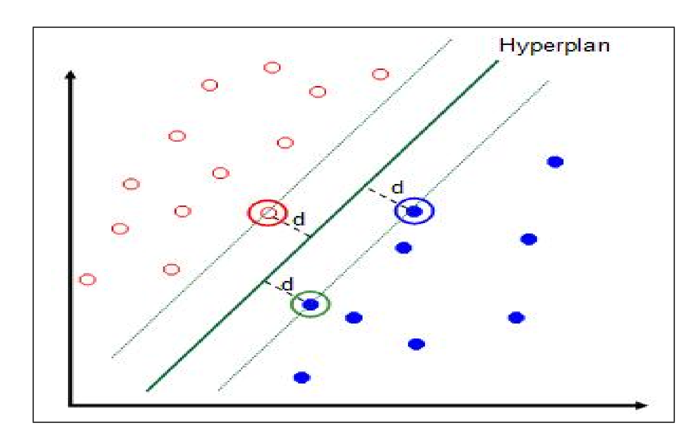


FIGURE 3.1 – Exemple de classification de deux types d'objets. L'hyperplan sépare les deux types de classes avec une marge de d.

$$P(h/D) = \frac{P(D/h) \times P(h)}{P(D)}$$

Avec $P(h/D)$: Probabilité de l'hypothèse h sachant D (probabilité a posteriori).
$P(h)$: Probabilité de h soit vérifiée indépendamment des données D (probabilité a priori).
$P(D)$: Probabilité d'observer des données D indépendamment de h. $P(D/h)$: Probabilité d'observer des données D sachant que h est vérifiée.
Ce théorème repose sur l'hypothèse que des solutions recherchées peuvent être trouvées à partir de distributions de probabilité dans les hypothèses et dans les données. Cette hypothèse d'indépendance ne reflète pas la réalité d'où l'appellation *naïf*. La classe la plus proche d'un nouvel objet est déterminée en combinant les prédictions de toutes les prédictions de toutes les hypothèses en les pondérant par leur probabilité a priori. Pour un ensemble de classes C et une instance spécifiée par un ensemble d'attributs A, la valeur de classification bayésienne naïve c est définie comme suit :

$$c = argmaxP(c_j)c_jP(a_i/cj)$$

Cette méthode de classification s'est montrée moins performante pour des tâches de classifications de textes [164].
Les arbres de décision Les arbres de décision sont les plus populaires des méthodes d'apprentissage, leur première implémentation remonte aux années 1970 [18]. Le terme « arbre de décision »recouvre plusieurs types d'arbres en fonction de l'objectif. On parle d'arbre de classification lorsqu'il s'agit de prédire la classe d'appartenance d'un objet, d'arbre de régression lorsque le résultat est de prédire une valeur numérique. CART (*Classification And Regression Trees*) est un type qui réunit les deux. Les principaux algorithmes des arbres de décision sont ID3 [134], C4.5 et C5.0 [135]. Comme toute méthode d'apprentissage supervisée, les arbres de décision ont besoin d'exemples d'objets déjà classés. Ces exemples sont représentés sous

forme d'attributs/valeurs. Si la tâche considérée est la catégorisation de textes, les exemples sont sous la forme de couples (texte i, catégorie k).
Le principe de construction d'un arbre de décision est assez simple. Il s'agit de déterminer les règles (appelées aussi questions ou tests) qui à chaque branche de l'arbre permettent de subdiviser l'ensemble de données en deux sous-ensembles plus homogènes. Ainsi la tâche de classification est binaire. Pour la catégorisation de textes, l'objectif sera de déterminer les règles (termes) qui permettent de subdiviser ces textes en fonction des attributs communs. La plupart des algorithmes d'arbre de décision arrêtent de subdiviser lorsque :

- La catégorie ne contient qu'un seul élément.
- Tous les éléments d'une catégorie ont les mêmes caractéristiques, donc la condition d'homogénéité est remplie.
- A la prochaine subdivision, l'amélioration attendue est si petite qu'elle ne justifie pas l'effort de subdivision.

Les règles sont souvent formulées sous la forme « *Si ... alors ...* ». Certains algorithmes disposent d'heuristiques pour déterminer ces règles. L'algorithme CART essaie toutes les règles. Il sélectionne ensuite la meilleure règle qui subdivise les données en deux ensembles en se basant sur la mesure de l'entropie (mesure d'incertitude associé au résultat d'un tirage aléatoire). L'algorithme ID3 emploie la mesure de *gain d'information* pour déterminer les règles de subdivision de l'arbre. Cette mesure repose sur la mesure de l'entropie :

$$E(S) = -p/Nlog2p/N - n/Nlog2n/N$$

Avec $E = entropie$, $S = des\ exemples\ de\ taille\ N$, $p = nombres\ d'exemples\ positifs$, n est le nombre d'exemple négatifs dans l'ensemble S des N.
Plus la valeur de l'entropie est petite, meilleure est la qualité des règles de subdivision et par conséquent plus homogènes sont les catégories obtenues.

3.4 Conclusion

Quelle que soit la méthode d'analyse de données, les résultats dépendent pour beaucoup de multiples paramètres : la taille du corpus, le choix de l'unité textuelle de représentation (mots, n-grammes, syntagmes nominaux, termes), les prétraitements effectués (lemmatisation, retrait des mots vides, élimination des mots très fréquents), le nombre d'itérations de l'algorithme, le nombre de classes à former, etc.
Quelle méthode de fouille de données utiliser reste une question largement ouverte, bien que certaines méthodes semblent être indiquées pour certains types de tâches : les classifieurs SVM, arbre de décision ou k-NN semblent bien fonctionner sur la catégorisation de textes, les règles d'association pour la découverte des motifs inconnus et les réseaux bayésiens pour des problèmes de probabilités conditionnelles.
Rappelons que l'extraction d'information consiste, au sein d'un texte donné, à isoler

les différents segments pertinents au regard d'un besoin informationnel. Souvent l'information pertinente se présente autour d'un concept particulier du domaine traité qui nécessite alors une exploration conceptuelle (indexation conceptuelle) du texte pour la localiser. Les ontologies, comme ressource sémantique, sont utilisées pour aider à l'exploration du corpus. Dans la section suivante nous allons définir ce qu'est une ontologie, ses composants et les méthodes de la création d'ontologie qui vont servir dans notre travail pour mettre en œuvre une ontologie de compétences des entreprises.

Chapitre 4

Les ontologies

« C'est parce que les choses ont une essence que les mots ont un sens »Pierre Aubenque, le problème de l'être chez Aristote, 1994

4.1 Introduction

Le but d'un système d'extraction d'information est d'identifier les entités pertinentes dans un texte à l'aide de base de connaissances du domaine. Une Ontologie de références du domaine traité est nécessaire. L'ontologie a pour rôle de valider les entités identifiées dans le texte. Pour répondre à notre besoin d'information, nous avons besoin d'une ontologie qui décrit le domaine des compétences des entreprises. C'est pourquoi dans ce chapitre on s'intéresse à comprendre ce qu'est une ontologie ? Quels sont ses constituants et quelles méthodes d'ingénierie utiliser pour la construire.

4.2 Définitions des ontologies

Ontologie PHILO : *Partie de la métaphysique qui s'applique à l'être en tant qu'être, indépendamment de ses déterminations particulières* (Le Petit Robert).
Taxinomie 1. DIDACT. *Etude théorique des bases, des lois, des règles, des principes, d'une classification. 2. Classification d'éléments* (Le Petit Robert).
Ontologie INGENIERIE DES CONNAISSANCES. *Ensemble des objets reconnus comme existant dans le domaine. Construire une ontologie c'est aussi décider de la manière d'être et d'exister des objets.*
"Une compréhension partagée d'un domaine donné" [70].
"L'ontologie est une spécification formelle d'une conceptualisation partagée" [23].
Une ontologie est une spécification formelle d'une conceptualisation d'un domaine. Elle est formée par des concepts et des relations. Elle est utilisée pour renvoyer à des structures lexicales et sémantiques variées : les modèles entités-relations pour les bases de données ; les dictionnaires, le thesaurus pour l'informatique linguistique ; les index pour la RI ; les définitions de classes orientées objets pour l'ingénierie des systèmes, etc.
Une ontologie est composée de :

1. *Concepts* qui sont souvent représentés par des termes,
2. *Relations* entre ces concepts (*sous-classe-de* ou *partie-de*),

3. *Fonctions* qui sont des cas particuliers des relations dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-1 premiers,
4. *Axiomes* qui sont utilisés pour structurer des phrases toujours vraies,
5. *Instances* qui sont utilisées pour représenter les éléments.

Nous n'allons pas trop détailler la définition de ces 5 éléments constructifs de l'ontologie, mais nous allons insister sur la définition du premier élément (les concepts) : le terme concept est souvent utilisé comme se référant à toute notion, de l'idée au lexème, en passant par l'entité et la catégorie. Selon Medin [112], un concept est une idée qui inclut tout ce qui est caractéristiquement associé à elle. Ces caractéristiques ont été décrites comme des conditions nécessaires et suffisantes des attributs définis pour une catégorie. Parce que nous utilisons des caractéristiques nécessaires et suffisantes pour décrire les catégories, cette approche s'avère très économique et permet de produire une seule représentation pour chaque catégorie.
Les chercheurs et les concepteurs des ontologies classent les ontologies existantes selon le degré d'implication de leurs composants. Si une ontologie contient seulement les concepts et les relations entre les concepts, on parle d'ontologie moins formelle ou « *light-weight* », et si l'ontologie contient en plus des fonctions et les axiomes qui offrent une capacité plus étendue de raisonnement sur les concepts, on parle alors d'ontologie formelle ou « *heavy-weight* »[143]. En pratique, il existe très peu d'ontologies « *heavy-weight* »qui regroupent tous les composants. En effet, appliquer les raisonnements sur les axiomes à un large ensemble de concepts, devient vite compliqué voire même impossible. Les ontologies les plus vastes, qui sont utilisées actuellement à grande échelle, simplifient cette représentation. Elles reposent sur la définition des concepts et des relations entre ces concepts. Parmi les ontologies « *Light weight* », on peut citer : Gene Ontology (GO) [1], MeSH [2], UMLS (domaine médical)[20], WordNet [116], EuroWordNet [161].
Définir une ontologie est une tâche de modélisation menée à partir des textes ou des corpus textuels qui représentent des expressions linguistique des connaissances d'un domaine spécifique. La modélisation s'effectue en trois étapes qui correspondent à trois engagements [9] : un engagement sémantique, fixant le sens linguistique des concepts, un engagement ontologique fixant leur sens formel et enfin un engagement computationnel déterminant leur exploitation effective.

4.3 Rôle des ontologies

Les ontologies peuvent jouer divers rôles qui sont :
– Acquisition et représentation des connaissances.

1. The Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource, Nucleic Acids Research, 32, D258-D261.
2. MeSH pour Médical Subject Heading, est un thesaurus contenant un vocabulaire contrôlé du domaine médical et un ensemble riche de relations liant les différents termes. Il est utilisé pour indexer des articles et ouvrages traitant du domaine médical. On peut l'explorer sur : http ://www.nlm.nih.gov/mesh/MBrowser.html

- Recherche et extraction des connaissances : inférer la connaissance qui est pertinente face à la requête de l'utilisateur ;
- Partage et intégration des connaissances : intégration de différentes sources d'information ;
- Gestion des connaissances ;
- Simplification du dialogue homme-machine.

Sur le Web l'utilisation de plusieurs ontologies permet de définir des spécifications relatives à plusieurs domaines. Cela permet à la machine de comparer une information reçue à des connaissances afin d'en tirer un sens et de pouvoir les exploiter. Cette représentation des connaissances est faite à l'aide des ontologies. C'est de là qu'est né le web sémantique.
L'expression web sémantique, attribuée à Tim Berners Lee au sein du W3C, fait d'abord référence à la vision du « web de demain »comme un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés (W3C, 2005).

4.4 Construction automatique d'ontologie à partir du texte

Les approches d'extraction automatique des termes et des relations à partir du texte, dans le cadre d'une aide à la création automatique d'une ontologie sont nombreuses [40] [41] [150]. Dans [40] les auteurs présentent une deuxième version de l'outil Caméléon d'extraction des relations qui est basée sur l'utilisation des patrons lexicale dédiés à l'extraction des relations sémantiques. Cette nouvelle version du système Caméléon permet de prendre en compte des textes annotés syntaxiquement. L'extraction de relations à partir de texte intervient de deux manières complémentaires dans le processus d'ingénierie d'ontologie : elle sert à l'identification automatique de relations entre classes de concepts (peuplement d'ontologies), elle peut aussi servir à l'extraction de relations ou de propriétés entre les classes de concepts et participe à la construction d'ontologie.
D'autres travaux [50] [153] [125] se sont intéressés à l'enrichissement automatique des ontologies à partir du texte en utilisant des techniques de fouilles de données. Cet enrichissement s'effectue en trois étapes : (i) extraction de termes représentatifs dans un domaine spécialisé (ii) identification de relation lexicales entre les termes (iii) placement des nouveaux termes dans l'ontologie existante.

4.4.1 Outils de TAL pour la construction de RTO

La construction automatique de Ressources Termino-Ontologiques (RTO) à partir de texte se résume à deux fonctions primordiales [29] :

- *Acquisition de termes* : une première classe regroupe les outils dont la visée est l'extraction à partir du corpus analysé de candidats termes, c'est-à-dire

de mots ou groupes de mots susceptibles d'être retenus comme termes par un analyste, et de fournir des étiquettes de concepts. Ces outils diffèrent principalement quant au type de techniques mises en œuvre (syntaxique, statistique, autres).
- *Structuration de termes et regroupement conceptuel* : Les ressources termino-ontologiques se présentent rarement sous la forme d'une liste à plat. Des outils d'aide à la structuration d'ensembles de termes sont donc nécessaires. Dans cette classe, nous évoquons, d'une part, les outils de classification automatiques de termes et d'autre part des outils de repérage de relation.

4.4.1.1 Acquisition de termes

TERMINO [48] est une application pour l'acquisition automatique de termes. Elle se focalise sur le repérage des syntagmes nominaux qui sont les seules structures supposées produire des termes. Ces candidats termes extraits sont appelés « *synapsies* ». ACABIT extrait des candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé [47]. Il mêle des traitements linguistiques en analysant le corpus étiqueté pour extraire des séquences nominales et les ramener à des termes binaires et des filtres statistiques grâce auxquels les candidats termes binaires sont triés au moyen de mesures statistiques. SYNTEX [27] s'appuie essentiellement sur une analyse syntaxique afin dextraire la terminologie du domaine. La méthode consiste à extraire les syntagmes nominaux maximaux. Ces syntagmes sont alors décomposés en termes de *têtes* et d'*expansions* à l'aide de règles grammaticales. Les termes sont ensuite proposés sous forme de réseau organisé en fonction de critères syntaxiques. L'environnement SYMONTOS [160] propose des outils pour repérer des termes simples et complexes dans des textes et des critères pour décider de définir des concepts à partir de ces termes.

4.4.1.2 Structuration des termes et regroupement conceptuel

Les outils de structuration des termes et de regroupement conceptuel peuvent être classés en deux gammes :

1. La gamme des outils qui visent à rapprocher les termes à partir d'une analyse globale de l'ensemble de leurs occurrences. Ils touchent les termes fréquents. Ce sont des outils très utilisés dans les applications d'informatique documentaire ou d'extraction d'informations. Ils classifient les termes sur la base de leurs distributions de cooccurrence dans le texte. Par exemple les outils de la recherche d'information rapprochent les termes qui apparaissent fréquemment dans la même classe parce qu'ils possèdent sans doute une certaine proximité sémantique. Cette technique qui vise à rapprocher les termes qui ont des distributions syntaxiques analogues, est à la base de nombreux travaux [74] [5] [57] [99].

2. La gamme des outils de repérage de relations, qui travaillent au niveau des occurrences elles-mêmes. Ils détectent dans le corpus les mots ou contextes syntaxiques répertoriés comme susceptibles de « marquer »une telle relation entre deux éléments [80]. L'un des enjeux principaux avec ces outils concerne la généralité des relations et celle des marqueurs de relations. Un certain nombre de travaux en TAL et en IC (Ingénierie des Connaissances) sont consacrés à ce problème. Ils partent tous du même principe d'une recherche itérative alternée dans le corpus à la fois des marqueurs d'une relation donnée et des couples de termes qui entrent dans cette relation [40] [41] [50] [153].

4.5 Ingénierie d'ontologie

Plusieurs chercheurs [64] [72] [61] ont pu démontrer que le concept d'ontologie permet d'analyser et de traiter le savoir dans un domaine en modélisant ses concepts pertinents. L'analyse de l'état de l'art dans le domaine de l'acquisition des connaissances et de la construction d'ontologie montre la nécessité d'utilisation d'une méthode basée sur un processus général permettant de passer des données brutes à l'ontologie. Ce cadre méthodologique consiste en général en quatre étapes, relativement indépendantes, qui s'accompagnent d'un double mouvement, du linguistique au conceptuel et de l'informel vers le formel. L'enjeu est de passer de la forme linguistique des connaissances, tirées du corpus du domaine, à la forme logique permettant son exploitation informatique :

- Construction d'un corpus de documents : Il s'agit de rassembler un ensemble de documents en relation avec le domaine d'application traité. Ces documents peuvent être des manuels techniques, des ouvrages, des transcriptions d'interviews menées avec des spécialistes du domaine... Ce corpus de documents formé contient des expressions linguistiques et des termes du domaine qu'il faut analyser.
- Analyse linguistique ou statistique du corpus : c'est une analyse pour rechercher les données conceptuelles dans les textes en utilisant des méthodes et des outils de traitement de la langue naturelle (TAL).
- Normalisation sémantique et formation des concepts : c'est une étape qui consiste à associer aux termes une signification et un concept qui fasse abstraction des variations de sens liées aux différents contextes textuels. A ce stade l'ontologie constituée est informelle.
- Elaboration de l'ontologie computationnelle : dans cette étape on traduit l'ontologie obtenue à l'étape précédente en une ontologie computationnelle spécifiée c.à.d dans un langage de programmation doté de services inférentiels. C'est une représentation formelle de l'ontologie.

A partir de ce cadre méthodologique général, qui montre que la construction d'une ontologie consiste à établir un ensemble de primitives dont la signification établit un modèle de la réalité, dérivent plusieurs méthodologies de construction d'ontologie qui font l'object de la section suivante.

4.5.1 Méthode d'ingénierie des ontologies

La construction d'ontologie est un processus complexe qui nécessite la mise en place de nombreux principes et critères. Du fait de cette complexité et de cette difficulté de construction, il n'existe pas encore de contribution sur les meilleures méthodes et pratiques lors du processus de développement d'une ontologie. Dans la littérature, plusieurs écrits sont disponibles pour cette problématique [117] [143] [133].
Si on considère qu'une méthodologie est l'ensemble des principes de construction appliqués avec succès par un auteur dans la construction d'ontologies, [113] a pu dénombrer un total de trente trois méthodologies existantes. Ces méthodologies sont analysées selon le type du processus de construction : à partir du début, par intégration ou fusion avec d'autres ontologies, par re-ingénierie, par construction collaborative ou par évaluation des ontologies construites.
Dans la perspective de pouvoir identifier une méthodologie de construction d'ontologie qui répond à notre besoin, nous allons présenter quelques méthodes qui nous semblent les plus proches de nos directives et les principales dans un processus de construction d'ontologie. Nous tenons aussi à signaler que nous décrivons pour chaque méthodologie les procédures de travail, les étapes qui décrivent le pourquoi et le comment de la conceptualisation, puis l'artefact construit.

4.5.1.1 METHONTOLOY

C'est une méthodologie qui a été développée au sein du groupe d'ontologie à l'université polytechnique de Madrid. Elle est liée au travaux de *software development process* [1] et *knowledge engineering methodologies* [64] [163]. Cette méthodologie est basée sur : l'identification du processus de développement (spécification, conceptualisation, formalisation, implémentation, maintenance), le cycle de vie basé sur l'évolution de prototypes et les techniques de gestion de projet (planification, assurance qualité) et des activités de support (intégration, évaluation, documentation).
L'activité de conceptualisation organise et convertit une perception informelle du domaine. Une fois que le modèle conceptuel construit, METHONTOLOGY propose de le transformer en un modèle formel qui va être implémenté. En suivant cette méthodologie, le constructeur d'ontologie doit effectuer les tâches suivantes :
Tâche 1 : construire un glossaire des termes de l'ontologie, leurs définitions en langue naturelle, leurs synonymes et acronymes.
Tâche 2 : construire la taxonomie des concepts pour les classifier.
Tâche 3 : construire les diagrammes de relation binaire ad hoc pour identifier les relations entre les concepts.
Tâche 4 : construire un dictionnaire des concepts, qui inclut principalement les instances de chaque concept, leurs instances et attributs et leurs relations ad hoc.
Tâche 5 : décrire en détail les relations binaires qui apparaissent dans les diagrammes de relations et les diagrammes de concepts.
Tâche 6 : décrire en détail chaque instance attribut qui apparaît dans le dictionnaire

de concept.
Tâche 7 : décrire en détail chaque classe d'attribut qui apparaît dans le dictionnaire de concepts.
Tâche 8 : décrire en détail chaque constante et produire une table de constantes. Les constantes spécifient l'information relative au domaine de connaissance, prennent toujours la même valeur et sont normalement utilisées dans les formulaires.
Tâche 9 : définir les axiomes formels.
Tâche 10 : définir les règles.
Tâche 11 : définir les instances.

4.5.1.2 Méthode de Uschold et King

Uschold et King [157] ont proposé une méthode de construction d'ontologie basée sur l'expérience acquise lors du développement de l'ontologie *The entreprise ontology*. Cette méthodologie est basée sur les étapes de construction suivantes :
- Identification des objectifs et du contexte de l'ontologie : clarifier le pourquoi de la construction de l'ontologie et les utilisations prévues.
- Construction d'ontologie : cette étape est divisée en trois activités :

Activité 1 : capture de l'ontologie : identification des concepts et des relations clés, pour produire en langage naturel les définitions précises et non ambigües de ces concepts. Pour réaliser cette activité Uschold et King proposent trois approches :

1. Approche descendante : partir de concepts abstraits que l'on spécialise en concepts plus spécifiques.
2. Approche ascendante : partir de tous les concepts spécifiques que l'on généralise en concepts abstraits.
3. Approche intermédiaire : les concepts se structurent autour de concepts intermédiaires, ni trop généraux, ni trop spécifiques.

Activité 2 : codage de l'ontologie : cette activité inclut la représentation explicite de la conceptualisation (classe, entité, relation) et l'écriture du code dans un langage formel (Prolog, OIL, OWL...)
Activité 3 : intégration d'ontologies existantes : Evaluation et documentation de l'ontologie

4.5.1.3 La méthodologie On-To-Knowledge (OTK)

On-To-Knowledge [143] est un projet qui vise à appliquer les ontologies aux informations et ressources textuelles disponibles sur l'internet, extranet et internet. Cette méthodologie propose de construire une ontologie très dépendante de l'application, qui tient compte du cycle de vie et de la future utilisation de l'ontologie.
La méthode On-To-Knowledge propose les étapes suivantes :
Etape 1 : Etude de faisabilité : c'est une étape qui adopte l'étude de faisabilité. Elle est appliquée sur l'application entière et sert de base à l'étape suivante.
Etape 2 : *Kickoff* : c'est une étape qui consiste à décrire les spécifications des

besoins de l'ontologie :

- Le domaine (contexte) et l'objectif de l'ontologie.
- Les directives de conception (les conventions de nommage ...)
- Les sources de connaissance et d'informations valables (livres, magazines, interviews...)
- Les utilisateurs potentiels et les cas d'utilisations.

Etape 3 : Raffinement : cette étape consiste à produire une application conformément aux spécifications données à l'étape de Kickoff. Cette étape est divisée en deux activités :

- Mise à jour des connaissances avec les experts du domaine : la première version de l'ontologie obtenue à l'étape précédente est raffinée au moyen d'interactions avec les experts du domaine.
- Formalisation : c'est l'implémentation de l'ontologie dans un langage d'ontologie. On-To-Knowledge recommande l'éditeur d'ontologie OntoEdit qui offre la possibilité de générer automatiquement le code d'ontologie dans plusieurs langages.

Etape 4 : Evaluation : cette étape sert à prouver l'utilité du développement de l'ontologie et les applications associées. Elle comporte deux activités :

- Contrôler si l'ontologie satisfait les spécifications (besoins).
- Tester et évaluer l'ontologie dans le cadre de son environnement d'application. Plusieurs allers retours sont nécessaires avant d'atteindre le niveau de satisfaction souhaité.

Etape 5 : La maintenance : Préciser comment s'effectue la maintenance. On-To-Knowledge propose que la maintenance de l'ontologie soit effectuée comme une partie de l'application.

4.5.1.4 La méthode SENSUS

C'est une méthode qui propose de construire une ontologie du domaine à partir d'une plus grande ontologie, l'ontologie SENSUS [155]. La méthode propose de relier les termes spécifiques du domaine à cette ontologie et d'élager dans SENSUS, les termes qui ne relèvent pas de la nouvelle ontologie qu'on souhaite construire. Durant le processus de construction, les étapes suivantes sont recommandées :
Etape 1 : identifier les termes clés du domaine.
Etape 2 : relier manuellement les termes clés à SENSUS
Etape 3 : inclure tous les concepts qui se trouvent sur le chemin depuis le terme clé jusqu'à la racine de SENSUS.
Etape 4 : ajouter manuellement les termes utiles pour le domaine et qui ne sont pas encore apparus. Reboucler sur les étapes 2 et 3 pour inclure les concepts sur le chemin, les nouveaux concepts jusqu'à la racine de SENSUS.
Etape 5 : ajouter le sous arbre entier.

4.5.1.5 La méthode ARCHONTE

La méthode ARCHONTE (ARCHitecture for ONTological Elaborating) proposée par Bachimont [7] [9] pour construire des ontologies s'appuie sur la sémantique différentielle. La composition d'une ontologie comporte trois étapes (figure 4.1 :

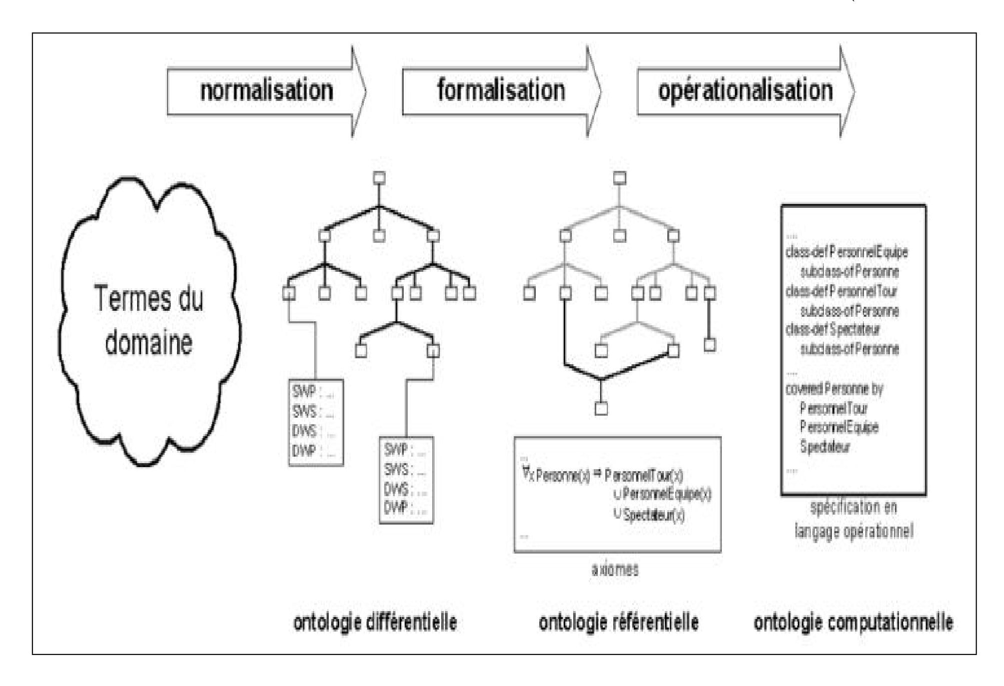


FIGURE 4.1 – La méthode ARCHONTE

1. Choisir les termes pertinents du domaine et normaliser leur sens puis justifier la place de chaque concept dans la hiérarchie ontologique en précisant les relations de similarités et de différences que chaque concept entretient avec ses concepts frères et son concept père.
2. Formaliser les connaissances, ce qui implique par exemple d'ajouter des propriétés à des concepts, des axiomes, de contraindre les domaines d'une relation.
3. L'opérationnalisation dans un langage de représentation des connaissances.

Comme le montre la figure ci dessus, la méthode ARCHONTE comporte initialement trois étapes : la normalisation, la formation et l'opérationnalisation. L'idée principale de cette méthode est de proposer à partir des expressions linguistiques une ontologie référentielle qui s'opérationnalisera en une ontologie computationnelle. B. Bachimont propose de contraindre l'ingénieur des connaissances à un "engagement sémantique", c'est-à-dire à expliciter clairement le sens de chacun des concepts de l'ontologie, en introduisant une "normalisation sémantique". *"Les primitives nécessaires à la représentation des connaissances doivent être modélisées à partir des données empiriques dont on dispose, à savoir l'expression linguistique des connaissances. Le travail de modélisation doit s'effectuer à partir de documents attestés dans la pratique d'un domaine et rassemblés en un corpus. Le corpus est constitué de documents produits dans le contexte où le problème à résoudre se pose"* [7].

La construction d'une ontologie en suivant cette méthodologie, consiste à établir un ensemble de primitives dont la signification sera établie relativement à un modèle de la réalité. Ces primitives et ces trois étapes qui constituent le cœur de cette méthodologie sont détaillées ci-après :

La conceptualisation : La conceptualisation consiste à identifier, dans un corpus, les connaissances du domaine. La découverte des connaissances d'un domaine peut s'appuyer à la fois sur l'analyse de documents et sur l'interview d'experts du domaine. De même, l'analyse informelle des textes doit être doublée par une analyse automatique qui permet de détecter les termes et structures sémantiques (définition, règle) présentes dans le corpus. Certaines connaissances implicitement utilisées dans le domaine ne sont cependant jamais exprimées, ni dans le corpus, ni par l'expert car elles vont de soi pour tous. Un des points les plus délicats de la conceptualisation consiste donc à identifier ces connaissances.

L'ontologisation : Après la phase de conceptualisation, il convient de formaliser au cours de la phase d'ontologisation, le modèle conceptuel obtenu. Cinq critères permettent de guider le processus d'ontologie :

- La clarté et l'objectivité des définitions, qui doivent être indépendantes de tout choix d'implémentations.
- La cohérence (consistance logique) des axiomes ;
- L'extensibilité d'une ontologie, c'est-à-dire la possibilité de l'étendre sans modification ;
- La minimalité des postulats d'encodage, ce qui assure une bonne portabilité ;
- La minimalité du vocabulaire, l'expressivité maximum de chaque terme.

De même il faut bien voir que l'ontologisation est une traduction dans un certain formalisme de connaissances ; le respect de la sémantique du domaine doit être assuré par un engagement ontologique, notion proposée initialement par T.Gruber comme un critère pour utiliser une spécification partagée d'un vocabulaire [67]. Pour T.Gruber, un engagement ontologique est une garantie de cohérence entre une ontologie et un domaine, mais pas une garantie de complétude de l'ontologie. N.Guarino [70] définit l'engagement ontologique comme une relation entre un langage logique et un ensemble des structures sémantiques. Plus précisément, le sens d'un concept est donné par son extension dans l'univers d'interprétation du langage. Ces engagements, sémantiques et ontologiques doivent être garantis par une structuration sémantique des connaissances. Cette structuration est nécessaire pour combler le fossé formel entre les connaissances et le formalisme utilisé pour les représenter en machine. Une fois le modèle conceptuel structuré, il faut le traduire dans un langage semi-formel de représentation d'ontologie. Parmi les langages de représentation développés au niveau conceptuel, trois grands modèles sont distingués :

- Les langages à base de frame.
- Les logiques de description.
- Les modèles des graphes conceptuels.

Quelques uns de ces langages, ou des langages utilisant ces modèles, sont déjà opérationnels et les ontologies exprimées dans ces formalismes peuvent être directement utilisées en machine. Dans les autres cas, une opérationnalisation de l'ontologie est

nécessaire.
L'opérationnalisation : La dernière phase de construction de l'ontologie consiste à outiller une ontologie pour permettre à une machine de manipuler des connaissances du domaine. La machine doit donc pouvoir utiliser des mécanismes opérant sur les représentations de l'ontologie. Enfin l'ontologie opérationnalisée est intégrée en machine au sein d'un système manipulant le modèle de connaissances via le langage opérationnel choisi.
Comme les ontologies doivent être considérées comme des objets techniques évolutifs et possédant un cycle de vie qui nécessite d'être spécifié, [62] a proposé un cycle de vie inspiré du génie logiciel, qui inclut les étapes de la construction de l'ontologie. Ce cycle de vie comprend une étape initiale d'évaluation des besoins, une étape de construction, une étape de diffusion et une étape d'utilisation. Après chaque utilisation significative de l'ontologie, les besoins sont réévalués et l'ontologie peut être étendue et, si nécessaire, en partie reconstruite.

4.6 Conclusion

La présence d'un concept dans le texte n'est pas une condition suffisante pour marquer l'information pertinente. Des phénomènes linguistiques peuvent dériver le sens des mots et le même mot peut présenter deux sens différents selon le contexte d'utilisation. La phrase « nous fabriquons des machines d'usinage »et la phrase « nos machines d'usinage sont fabriquées par nos partenaires »présentent le même concept « fabrication »sauf que ce n'est pas le même sens. Dans la phase d'extraction, ces genre d'ambigüité (voir une simple négation « nous ne fabriquons pas de machines d'usinage ») sont perturbants pour notre objectif final (la détection de la bonne compétence chez l'entreprise). Pour lever cette ambigüité contextuelle, nous avons recours à une analyse fine du texte en utilisant les outils et ressources du traitement automatique de la langue naturelle. C'est un quatrième domaine dont nous nous servons pour résoudre le problème posé.
Nous avons argumenté ci-dessus l'utilité des ontologies comme support à l'extraction d'information dans certains contextes. Nous montrons dans la suite que c'est particulièrement pertinent pour traiter le domaine des compétences des entreprises. Dans cette perspective, nous avons insisté sur les méthodes de construction et d'ingénierie d'ontologie nécessaire à notre travail.

Chapitre 5

Traitement automatique de la langue

5.1 Introduction

Le corpus utilisé pour l'extraction de l'information dans notre travail est constitué de sites web des entreprises. Un site web d'une entreprise est un document mal structuré, contenant des données hétérogènes (publicité sur les produits, fondement de l'entreprise, employés, activités...). Les problèmes posés par les caractéristiques de ce corpus sont nombreuses, que ce soit au niveau du pré-traitement ou de l'interrogation. Face à ces problèmes, nous avons étudié des solutions spécifiques que pourrait apporter le Traitement Automatique de la Langue (TAL). Nous n'allons pas présenter en détail le domaine du TAL et ses enjeux, mais nous développons dans la suite les aspects de traitement automatique de la langue qui touchent à notre problématique : nous allons commencer par définir les différents niveaux d'analyse du langage (morpho-lexical, syntaxique, sémantique et pragmatique). Nous allons nous attarder sur un aspect important de ce domaine qui est l'extraction des schémas textuels en utilisant les patrons linguistiques. Dans une dernière section nous présentons le système *UNITEX* qui a été utilisé dans notre travail comme outil linguistique pour l'analyse du texte et l'extraction de l'information recherchée.

5.2 Analyse linguistique des textes

Les systèmes de traitements de l'information doivent fonctionner dans différents domaines de connaissances exprimés par des ressources textuelles pour pouvoir produire, diffuser, rechercher, exploiter et traduire les documents. C'est pour cela qu'ils ont besoin d'une analyse fine des textes pour bien préciser le sens des mots sans ambiguïté. Cette nécessité vient défendre le domaine du TAL (Traitement Automatique de la langue) comme clé pour une analyse linguistique fine et dépourvue d'ambigüité et pour une représentation du sens des mots du texte.
Les systèmes d'extraction d'information reposant sur les techniques de TAL doivent mettre en œuvre un traitement linguistique sur le texte à savoir la segmentation, l'analyse morphologique, la reconnaissance des entités nommés, la représentation sémantique des motifs extraits (si nécessaire), etc. les techniques de TAL se mobilisent conjointement avec des ressources spécialisées (lexiques, grammaires, dictionnaires, ontologies etc.) pour élaborer un système d'extraction d'information.

5.2.1 Les niveaux d'analyse linguistique

Pour comprendre un élément textuel (texte, phrase, proposition, mot ...), il faut combiner le sens des unités de taille inférieures. Le but d'une analyse linguistique est de montrer ce que sont les mots. Que signifient-ils ? Comment se combinent-ils pour former la phrase ? Et, par ailleurs, comment calculer le sens d'une unité plus grande ?
Dans notre cadre de travail, nous nous basons uniquement sur la langue écrite (l'analyse des sites web des entreprises) ce qui implique que les entités les plus petites que nous allons étudier sont les mots. En conséquence, l'analyse de notre corpus textuel (site web des entreprises) peut se rapporter à quatre niveaux :
L'analyse morpho lexicale : qui se préoccupe de la structure des mots.
L'analyse syntaxique : étudie les règles liant les unités linguistiques entre elles et contrôle la bonne formation de la phrase.
L'analyse sémantique : qui s'intéresse au sens des phrases considérés individuellement.
L'analyse pragmatique : définit un contexte autour de chaque phrase.

5.2.1.1 L'analyse morpho-lexicale

Elle a comme objectif d'identifier les mots du texte (simples, composés, noms propres, abréviations) et leurs traits (genres, nombre, mode, temps etc.). Elle représente également l'étude des règles de combinaison des morphèmes (unités minimales de sens). En pratique dans le cadre de traitement automatique de la langue naturelle, l'analyse morpho lexicale consiste en une succession des étapes suivantes :

1. Segmentation : découpage du texte en phrases puis en mots distincts (*Tokenisation*).
2. Lemmatisation : elle consiste à associer un lemme à chaque mot du texte. c'est la forme canonique d'un mot qui regroupe les différentes formes que peut revêtir un mot : le genre, le nombre, la lexion, etc.
3. Etiquetage : identifier la bonne catégorie morpho-syntaxique (nom, verbe, adjectif, etc.) des mots selon le contexte.

Chacune de ces trois étapes est très importante car elle conditionne le contexte du mot. En effet un même mot qui s'écrit de la même façon peut avoir plusieurs interprétations différentes. Considérons par exemple les deux phrases suivantes :
Phrase 1 : *Une entreprise a le produit.*
Phrase 2 : *Elle l'a produit.*
Une analyse morpho-syntaxique avec l'outil TreeTagger [1] donne le résultat suivant

1. http ://perl.linguistes.free.fr/ telechargements.html
C'est un outil pour l'annotation grammaticale de données textuelles, par l'association à chacun des mots partie du discours son genre : noms, verbes, adj, etc et son lemme. Cet outil a été développé par HELMUT SCHMID dans le cadre du projet " TC " à l'institut de Linguistique informatique de l'Université de Stuttgart. TreeTagger a été utilisé avec succès pour différentes langues : allemand, anglais, français, italien, chinois. Il est fondé sur un algorithme d'arbre de décision pour effectuer l'analyse grammaticale.

(table 5.1 : La même forme (produit) qui s'écrit de la même façon dans les deux

Mot	Catégorie Grammaticale	Lemme
Une	DET :ART	un
entreprise	NOM	entreprise
a	VER :pres	avoir
le	PRO :PER	la/le
produit	NOM	produit
.	SENT	.
Elle	PRO :PER	la/le
l'	PRO :PER	la/le
a	VER :pres	avoir
produit	VER :pper	produire
.	SENT	.

Table 5.1 – Analyse Morphosyntaxique des deux phrases "Une entreprise a le produit" et "Elle l'a produit"

phrases donne lieu à deux interprétations grammaticales différentes. Dans la première phrase, il a indiqué le mot produit comme un (nom). Tandis que dans la deuxième phrase il a indiqué le mot produit comme un verbe (ver :pper).

5.2.1.2 Analyse syntaxique

L'objectif de cette étape est de structurer une chaîne d'unités lexicale en unités syntaxiques (syntagmes) et de déterminer comment les mots se combinent pour former des syntagmes puis des propositions et enfin des phrases correctes. C'est aussi la procédure permettant de décider si une phrase appartient ou non à un langage. Souvent le résultat de l'analyse syntaxique est représenté sous une forme hiérarchique (figure 5.1 : Pour tester si une phrase est correcte, on doit trouver une application des règles d'une grammaire qui l'engendre. Une grammaire est composée de :

- Un vocabulaire terminal, l'alphabet sur lequel est défini le langage.
- Un vocabulaire non terminal qui n'apparait pas dans les mots générés. Un symbole non terminal désigne une catégorie syntaxique.
- Un ensemble des règles de réécriture ou de production.
- Un symbole de départ. C'est à partir de ce symbole non terminal que l'on commencera la génération des mots au moyen des règles de la grammaire.

Par exemple la grammaire suivante valide la phrase *une entreprise a le produit* :
$S \longrightarrow NP, NV$
$NP \longrightarrow DET, N$
$VP \longrightarrow V, NP$
$DET \longrightarrow une$
$N \longrightarrow entreprise$
$V \longrightarrow a$

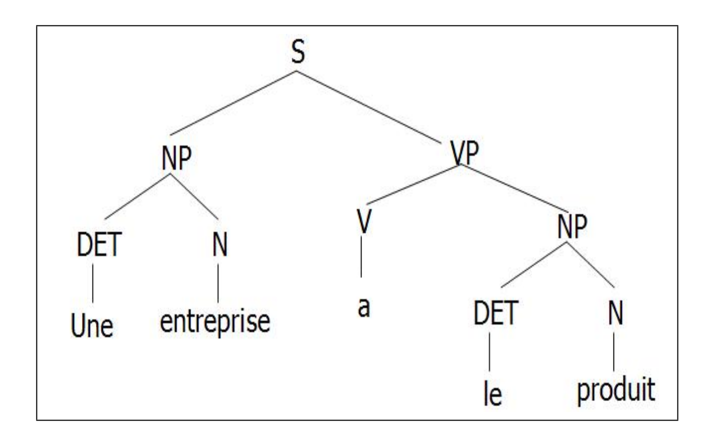


FIGURE 5.1 – Arbre syntaxique de la phrase "une entreprise a le produit" Avec S : sentence (phrase) ; NP : noun phrase (syntagme nominal) ; VP : verbal phrase (syntagme verbal) ; DET : déterminant ; N :nom ; V : verbe

$DET \longrightarrow le$
$N \longrightarrow produit$

5.2.1.3 Analyse sémantique

Cette étape essaie de donner un sens aux phrases du texte. Dans cette phase, les phrases sont traitées de manière isolée. Pour déterminer le sens d'une phrase, une première étape va se préoccuper du sens de chacun des mots constituant la phrase. Ensuite à l'aide des informations fournies par l'analyse syntaxique, le sens complet de la phrase pourra être déduit grâce à la connaissance des relations existant entre les mots. Pour ce faire une représentation du sens est nécessaire :
Représentation logique du sens
Il est possible de symboliser le sens d'un énoncé par une représentation logique à l'aide de prédicats possédant une syntaxe simple et dépourvue d'ambigüité ; ainsi la phrase *l'entreprise a le produit* peut être représenté comme suit :
$\exists x \exists y, entreprise(x) \wedge produit(y) \wedge avoir(x, y)$
De ce point de vue, déterminer la signification d'une phrase P d'une langue revient à établir les conditions de vérité de P dans l'ensemble des mondes possibles. Cette méthode a été introduite par Richard Montague en 1974 pour analyser un fragment de l'anglais. Un des principes gouvernant la grammaire de Montague (MG) est le principe de compositionnalité : à chaque règle syntaxique correspond une règle sémantique. Son analyse procède de la façon suivante : chaque phrase de la langue naturelle est traduite en une formule logique, toujours selon le parallélisme entre la syntaxique et la sémantique. Cette représentation logique de la phrase est ensuite évaluée dans l'ensemble des mondes possibles. En effet pour déterminer la signification d'une telle expression complexe dans une telle langue, on doit passer par la dérivation syntaxique (arbre). Malheureusement, la MG se heurte à certains problèmes d'interprétations des pronoms au-delà des limites de la phrase,

et en particulier, au problème des relations anaphoriques entre les pronoms et les descriptions définies. Pour cette raison au début des années 80, certains travaux ont cherché des voies alternatives à l'approche montagovienne, parmi lesquelles on trouve la théorie des représentations discursives (DRT).

Représentation avec la DRT

La DRT est une théorie de représentation du discours introduit par Kamp [90] qui traite dynamiquement les enchaînements de phrases à l'intérieur d'un discours et représente les phrases qui ne pouvaient être traduites par la logique des prédicats du 1er ordre. La DRT permet une représentation systématique et compositionnelle du discours. Elle traite la représentation de phénomènes linguistiques courants mais complexes comme la résolution d'anaphores, les phrases conditionnelles et l'emploi de quantificateurs.

Dans ce qui suit, on étudiera un exemple qui traite le problème de résolution de l'anaphore par les référents accessibles dans la Structures de Représentation du Discours (DRS). Si on considère la phrase "*L'entreprise fabrique des roulements. Elle a une bonne réputation*", sa représentation donne :

x, y, z
entreprise (x) roulements (y) fabrique (x, y) z = ? avoir-une-bonne-réputation (z)

Par résolution anaphorique on obtient :

x, y, z
entreprise (x) roulements (y) fabrique (x, y) z = x avoir-une-bonne-réputation (z)

Ces différents phénomènes et formalismes d'interprétation et de représentation de la sémantique n'ayant que peu de répercussions sur notre travail, nous ne nous attardons pas beaucoup sur ces questions.

5.2.1.4 Analyse pragmatique

Pour la bonne compréhension d'un texte, un lecteur a besoin de connaître un certain nombre d'éléments qui ne sont pas exprimés explicitement dans le texte : connaissances relatives à la culture générale, au sujet abordé, etc. L'enchainement

des étapes précédentes conduit parfois à des ambigüités qu'il est possible de supprimer en utilisant l'analyse pragmatique. Cette dernière, permet d'étudier le lien entre les unités linguistiques et leur contexte. Ainsi la phrase *Là, tu tournes à droite* ne peut avoir un sens complet et correct que si le lecteur possède une vision pragmatique claire sur la position de la personne en question.

5.2.2 Relations linguistiques et patrons

5.2.2.1 Acquisition des termes et des relations

Les mots dans un texte sont dépendant les uns des autres, ils sont employés dans un discours où des relations sémantiques peuvent être exprimées à travers une série de motifs morphologiques, lexicaux et syntaxiques. Les outils d'aide à la construction des relations terminologiques à partir de corpus textuel ont connu un essor important. *Syntex* [28] prend en entrée un corpus de phrases étiquetées et calcule pour chaque phrase les relations de dépendance syntaxique entre les mots (sujet, complément d'objet, complément prépositionnel, etc). L'analyse et l'extraction des relations terminologiques à partir du texte sont aussi utilisées pour la construction des ontologies [6] [42] [118]. Ces travaux partent du principe que les textes expriment également des informations sur les relations sémantiques que les termes entretiennent entre eux, et si on considère que les termes représentent l'ontologie à construire, les relations qu'ils entretiennent peuvent être considérées comme le reflet des relations conceptuelles de l'ontologie à construire.
Dans la plupart des cas, l'extraction terminologiques s'intéresse essentiellement à l'identification de syntagmes nominaux (mots isolés pour les noms "N" et les termes simples, schémas de type *N de N* ou *N à N N ADJ*, etc. pour les termes complexes, les approches d'extraction des relations reposent sur l'utilisation de patrons linguistiques selon lesquels une relation sémantique entre termes telle que l'hyperonymie ou la méronymie peut être abstraite dans un schéma linguistique qui décrit toutes les réalisations langagières associées.

5.2.2.2 Les patrons linguistiques

Les patrons linguistiques sont le résultat d'une observation de la réalisation d'une relation sémantique dans le texte afin d'en schématiser le contexte lexical et syntaxique. Cette schématisation constitue un patron lexico-syntaxique et permet d'extraire des couples de mots vérifiant cette relation à partir d'un corpus textuel.
Un patron linguistique est défini comme une « *forme linguistique faisant partie de catégories prédéfinies (grammaticales, lexicales, syntaxiques ou sémantiques) dont l'interprétation définit régulièrement le même rapport de sens entre les termes* »[75].
D'une façon plus élaborée un patron lexico-syntaxique identifie la relation recherchée plus précisément en définissant également des contraintes syntaxiques ou typographiques sur le contexte des termes [65].
En linguistique, les approches par patrons sont utilisées pour associer des régularités structurelles à des informations sémantiques. Hearst [80], est la première à utiliser

les patrons dans le contexte de l'extraction d'information. Elle a proposé des ensembles des patrons lexico-syntaxiques qui sont facilement repérables dans le texte et qui apparaissent fréquement dont le but de reconnaître certaines relation lexicales sans ambigüités. Hearst montre à partir de l'exemple de la phrase : "*The bow lute, such as the Bambara ndang, is plucked*" sans même savoir ce que sont un Bambara ndang et un bow lute, le lecteur est capable d'indiquer qu'un Bambara ndang est une sorte de bow lute. Dans cette phrase la relation d'hyponymie peut être encodée par la construction linguistique suivante : « *un terme suivi par such as et un autre terme* ». Elle est abstraite au sein du patron suivant :
X *such as* Y, où X et Y sont des syntagmes nominaux.
Les patrons linguistiques sont utilisés aussi pour l'enrichissement des ontologies [40] [121]. L'objectif est d'exploiter la projection des patrons sur un corpus textuel pour enrichir une ontologie existante. Les patrons ont pour fonctions d'extraire les relations entre les concepts présents dans cette ontologie ou d'extraire des nouveaux concepts qui seront ajoutés dans cette ontologie. Dans une perspective d'automatisation de l'enrichissement de l'ontologie, les auteurs effectuent un apprentissage de patrons lexico-syntaxique.
La définition manuelle des patrons à partir d'un corpus textuel est une tâche fastidieuse. Ceci a poussé les chercheurs à proposer des méthodes d'acquisition semi automatique. En général ces approches peuvent être décrites dans quatre étapes :

1. Normalisation du corpus ;
2. Filtrage des phrases pertinentes ;
3. Identification des exemples représentatifs ;
4. Génération de variantes à partir de patrons initiaux.

Afin de faciliter l'acquisition de patrons, le traitement du corpus est précédé d'une étape de normalisation. Après cette étape, un filtrage de séquences pertinentes est effectué. Il s'agit à partir d'un ensemble de mots-clés fourni par l'utilisateur, de ne retenir que les phrases potentiellement pertinentes dans le corpus pour éviter à l'expert de lire de larges pans de textes inutiles pour la tâche. Ensuite l'utilisateur détermine, parmi les phrases filtrées, les syntagmes représentatifs. Sur cette tâche, seule l'expertise humaine est capable de déterminer et d'évaluer la pertinence d'un syntagme ou d'un patron [169]. Enfin, le module de génération de variantes est capable d'étendre la couverture du système en proposant des structures prédicatives sémantiquement équivalente.

5.3 Le système UNITEX

Dans la deuxième partie de ce mémoire, le travail linguistique pour l'extraction des informations sur les compétences des entreprises a été réalisé avec le système Unitex.
Unitex est une réimplémentation *open-source* du système *Intex* [151]. En marge des possibilités de ce dernier, il intègre de nouvelles fonctionnalités dont la prise en

compte d'Unicode qui permet, sans traitement préalable, l'analyse de langues telles que l'arabe ou le grec, qui n'utilisent pas un alphabet latin. De plus, la mise à disposition du code source rend possible une collaboration à très grande échelle qui lui confère une vitalité nécessairement plus importante. Ce sont ces deux considérations qui nous ont amené à privilégier l'utilisation d'*Unitex*. Ce logiciel est téléchargeable [2] depuis le site de l'Institut Gaspard-Monge, promoteur du logiciel.
Unitex offre un cadre de travail très intéressant et accessible au linguiste non-informaticien. Il permet la formalisation graphique des automates sous la forme de *grammaires locales* et offre un nombre considérable d'outils permettant leur application au texte. Nous renvoyons, pour une description complète de ces outils, au manuel d'Unitex [129].
L'application Unitex a été développée au Laboratoire d'Apprentissage Documentaire et Linguistique (LADL) sous la direction de M. Maurice Gross. C'est un ensemble de logiciels permettant de traiter des textes en langage naturel en utilisant des outils linguistiques comme AGLAE [126] et INTEX [151]. Unitex [128] [127] est un environnement de développement utilisé pour construire des descriptions de textes formalisées sous la forme de dictionnaires électroniques et de grammaires représentés par des graphes à nombre fini d'états, et de lexiques grammaires pour des textes de taille importante. Il fournit des outils pour décrire et représenter les morphologies flexionnelle et dérivationnelle, les variations terminologiques, le vocabulaire (les mots simples, les mots composés et les expressions figées), les phénomènes semi-figés (grammaires locales, les accords) et la syntaxe. Unitex transforme tous les objets traités (les textes, dictionnaires, grammaires) en transducteurs à nombre fini d'état. Un transducteur est un graphe qui représente un ensemble de séquences en entrée et leur associe un ensemble de séquences en sortie.

5.3.1 Les dictionnaires

Représentés sous le formalisme DELA (Dictionnaires Electroniques du LADL), les dictionnaires électroniques permettent de décrire les entrées lexicales simples et composées d'une langue en leur associant un lemme avec une série de codes grammaticaux, sémantiques et flexionnels. Ils ont été élaborés pour plusieurs langues comme le français, l'anglais, le grec, l'italien, l'allemand, l'espagnol, le thaïlandais, le coréen, le norvégien, le portegais. Il existe deux sortes de dictionnaires électroniques : les premiers sont le DELAF (DELA de forme fléchie) et le DELACF (DELA de forme composée fléchie). Ce sont les plus utilisés. Les deuxièmes, les dictionnaires de forme non fléchie, sont le DELAS (DELA de forme simple) et le DELAC (DELA de forme composée).
Pour chaque langue le dictionnaire DELAF liste toutes les formes fléchies et les associe au lemme. Prenons l'exemple suivant :
$entreprises, entreprise.N + z1 : fp$
La forme *entreprises* est associé au lemme *entreprise*. La lettre N signifie que c'est un nom. $z1$ indique qu'il s'agit d'un mot du langage courant. Le code flexionnel : fp

2. http ://www-igm.univ-mlv.fr/ũnitex/

représente le féminin pluriel.
Autres exemple :
$machines - outils, machine - outil.N + NN + Conc + z2 : fp$
$mcanique, .N + z1 : fs$
Le DELACF pour le français contient 250000 formes de noms composés, 8000 adverbes figés, 15000 formes figées utilisées avec le verbe être et 1600 conjonctions de subordination.

5.3.2 Les grammaires

De nombreuses études ont mis en évidence l'adéquation des automates aux problèmes linguistiques. Ainsi, une grammaire décrit des séquences de mots et produit des informations linguistiques (sur la structure syntaxique par exemple). Un dictionnaire représente les séquences de lettres et produit les informations lexicales associées. Le transducteur d'un texte représente les séquences de mots qui représentent chaque phrase et leur associe des informations lexicales ou syntaxiques des résultats produits par différentes analyses.
Les grammaires sont représentées au moyen de graphes que l'utilisateur peut créer et mettre à jour. L'application de dictionnaires à un texte consiste à construire l'union des transducteurs de chaque dictionnaire avec le transducteur du texte. Une

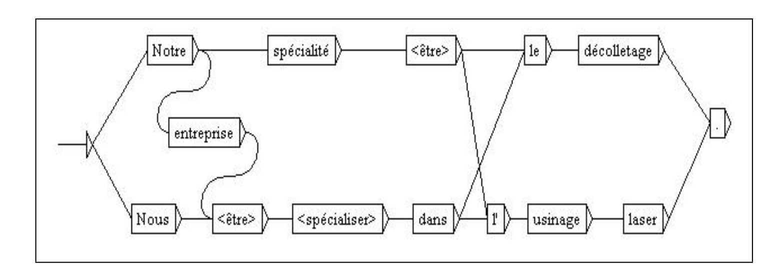


FIGURE 5.2 – Exemple d'une grammaire locale

grammaire locale est une représentation par automate de structures linguistique difficilement formalisables dans des tables de lexique-grammaire ou dans des dictionnaires électroniques. Les grammaires locales, représentées sous la forme de graphes, décrivent des éléments qui relèvent d'un même domaine syntaxique ou sémantique. Les descriptions linguistiques décrites sous la forme de grammaires locales sont utilisées pour une grande variété de traitements automatiques appliqués sur les corpus de texte. Ces grammaires locales sont un moyen puissant de représenter la plupart des phénomènes linguistiques. Ce sont des variantes des grammaires algébriques, également appelées grammaires hors-contexte.
Une des principales fonctionnalités d'Unitex est la recherche d'expressions dans des textes. Une fois que le texte a subi une opération de prétraitement (normalisation des formes non ambigües, découpage de texte en phrases) et que les dictionnaires électroniques ont été appliqués, on peut effectuer des recherches sur ces textes en leur appliquant les grammaires.

5.4 Conclusion

Nous envisageons, après cette étude des différents domaines sollicités, la mise au point d'un ensemble de méthodes et outils de recherche et d'extraction d'information répondant à notre besoin (détection des activités des entreprises et extraction des traces de leurs compétences). Partant d'un corpus textuel (sites web des entreprises), des ressources linguistiques du lexique-grammaire, nous voulons estimer la pertinence de la *phrase élémentaire ou de l'expression linguistique*.
Nous devons donc nécessairement nous positionner dans un contexte applicatif réel. Un tel objectif demande la prise en compte d'un grand nombre de phénomènes : l'ambiguïté lexicale, la complexité syntaxique, l'anaphore, etc. Dans la mesure où ces phénomènes peuvent poser problème, nous envisagerons des solutions afin, au final, d'obtenir un extracteur d'information fonctionnel.
Toutefois, parce que certains problèmes sortent du cadre précis de cette étude, les solutions que nous apporterons peuvent être partielles et doivent être considérées comme les prémices d'une réflexion plus importante.
Pratiquement, nous procèderons en deux temps. Nous commençons par présenter un premier système de détection des activités des entreprises basé essentiellement sur des outils et méthodes de la recherche d'information. En deuxième temps, nous présentons un deuxième système d'extraction des compétences des entreprises basé sur les techniques d'extraction d'information, une ontologie descriptive du domaine et les outils et méthodes de traitement automatique de la langue.

Conclusion Partie 1

L'objectif général de la thèse est de contribuer à une méthodologie de recherche et d'extraction d'informations et de proposer des outils d'aide à la constitution de réseaux d'entreprises dans un environnement ouvert où les organisations ne se connaissent pas et ont une information hétérogène publique et non restreinte. Scientifiquement cette thèse est positionnée principalement sur le domaine de la recherche et du traitement automatisé de l'information. Les problématiques de recherche d'information que nous visons interviennent dans un contexte de recherche de collaboration inter-entreprises. Le cœur de la thèse consiste notamment à mettre au point une méthode et des outils de traitement de l'information s'appuyant sur l'utilisation des ressources sémantiques externes telle que les ontologies, permettant d'utiliser des informations publiques disponibles sur des entreprises d'un territoire ou d'un domaine d'activités donné, afin de faire émerger des propositions opérationnelles de mise en réseau de ces entreprises pour répondre au besoin du marché.

L'enjeu de la thèse est de contribuer à une automatisation de la recherche d'informations caractérisant des entreprises, en vue d'appliquer les modèles formels d'aide à la décision qui visent à identifier des collaborations inter-entreprises. Ainsi, en se référant à un modèle mathématique existant visant à identifier des réseaux d'entreprises, l'objectif sera de mettre au point des dispositifs d'analyses des informations utiles à l'application du modèle.

La recherche d'information dans une finalité donnée, à partir de données fournies en formats hétérogènes et relevant du domaine publique ou privé, requiert d'utiliser des mécanismes avancés permettant de manipuler la syntaxe mais également la sémantique des informations. Dans cet objectif, la thèse s'appuiera ainsi sur les techniques de recherche d'information, d'extraction d'information, des ontologies et de traitement automatique de la langue. En effet ces travaux de recherche s'inscrivent dans le cadre de la mise en œuvre de système de recherche et d'extraction automatique d'information à partir du web. L'objectif est de proposer un environnement ouvert sur les informations des entreprises. Cet environnement est utilisé pour construire un outil d'aide à la décision pour faire émerger des propositions de collaboration inter-entreprises. Il utilise une représentation sémantique profonde de l'information en se basant sur les ontologies et les liens sémantiques.

Le modèle formel d'aide à la décision que nous visons à appliquer suite à nos travaux de recherche et d'extraction d'information se réfère à la théorie économique de coordination entre entreprises, ainsi qu'aux travaux sur la constitution des réseaux d'entreprises développés par [15] [31]. En effet, Benali cherche, pour construire un réseau d'entreprise collaboratif, deux informations essentielles à la coordination des entreprises : la complémentarité des activités et la similarité des compétences. Dans ses travaux, la recherche d'informations est faite manuellement via un questionnaire. Ainsi pour la complémentarité des activités, il construit un graphe entre les différentes entreprises (nœuds du graphe). La pondération des arcs du graphe

de complémentarité représente le degré de complémentarité entre deux entreprises. Pour obtenir ce degré de complémentarité il utilise deux éléments, qui sont le chiffre d'affaires (C.A.) en pourcentage de chaque activité (gamme, classe, ou famille de produits) de chaque entreprise du réseau, et l'influence que peut avoir chacune de ces activités sur les activités des autres entreprises. Ainsi synthétiquement, les étapes de sa méthode sont les suivants :

- Obtenir la gamme (famille) de produits par entreprise.
- Affecter les chiffres d'affaires (C.A.) en pourcentage par classe de produit.
- Construire la matrice des degrés d'influence DI_{ij} entre deux produits A_i et B_j , en évaluant quels produits de l'entreprise A réagissent à une variation d'un produit de l'entreprise B et à quel degré. La matrice est remplie après questionnement de chaque entreprise du réseau de la manière suivante : « Si un changement quelconque (plus rentable, moins coûteuse, augmentation ou diminution de la production) arrive dans la famille de produits B_j de l'entreprise B, quelle serait son influence sur la famille de produits A_i de votre entreprise ? ». La réponse est guidée, et pour chaque réponse, une personne affecte un nombre entre 0 et 3 (pas d'influence = 0, peu d'influence = 1, influence moyenne = 2, forte influence = 3).
- Calculer l'influence mutuelle (IM) pour chaque paire de familles (A_i, B_j). Cette influence n'est pas symétrique, c'est à dire que l'influence de A_i sur B_j est souvent différente de celle de B_j sur A_i : $IM_{ij} = C.A.\%deA_i \times C.A.\%deB_j \times Di_{ij}$
- Calcul du degré de complémentarité de l'entreprise A sur l'entreprise B par la somme des influences mutuelles divisée par 3 (pour normer l'échelle entre 0 et 1)

Une méthode de partitionnement de graphe est appliquée pour détecter les réseaux d'entreprises en fort lien de complémentarité.

Pour modéliser et quantifier les compétences, Benali utilise des concepts de la théorie des sous-ensembles flous. Des notions de distances et de proximités sont utilisées pour quantifier l'éloignement entre les différentes entreprises du réseau en termes de compétences. Il commence par définir un ensemble fini de compétences à partir d'un référentiel ou dictionnaire de compétences comme le ROME (Répertoire Opérationnel des Métiers et de l'Emploi). Ensuite, l'évaluation des compétences de chaque entreprise est effectuée directement à travers le questionnaire. Une matrice de distances inter-entreprises en termes de compétences est calculée en fonction du degré d'évaluation de chaque compétence. Enfin, pour identifier les entreprises proches en termes de compétences, il applique une analyse factorielle sur la matrice de distances suivie d'une classification hiérarchique.

Notre travail porte sur l'automatisation de la méthode de la collecte et de traitement de l'information nécessaire à l'application de la méthode de Benali en utilisant le web comme espace de recherche des données. C'est une recherche d'information spécialisée qui s'applique en génie industriel. Cette recherche s'effectue dans un domaine informationnel caractérisé par une information métier représentée par des textes qui ne suivent aucune structure standard ; la sémantique du vocabulaire utilisé est

très liée au domaine métier (vocabulaire contextualisé) ; la structure linguistique des textes est parfois absente ; l'ensemble de ces facteurs induit de forts risques d'ambiguïté. Ainsi, nous cherchons à montrer la valeur ajoutée de l'usage de ressources sémantiques propres au métier, ce qui se justifie par les performances finales du système de recherche et d'extraction d'information. Cet objectif de recherche induit deux problématiques majeures : une première vise à détecter le secteur d'activité de l'entreprise à partir de son site web en se basant sur une indexation contrôlée par un thesaurus décrivant tout les domaines d'activités des entreprises en France tel que le NAF (Nomenclature des Activités Françaises). Une deuxième problématique vise à extraire une information plus spécifique représentée par des fragments de texte (mot, expression, phrase) décrivant la compétence de l'entreprise. Cette dernière est rendue plus ardue par le fait qu'il n'existe pas des ressources sémantiques constituant des points de départs pour une telle analyse d'extraction d'information.
Aujourd'hui les résultats dans le domaine de la recherche et l'extraction d'information sont assez prometteurs, ce qui nous a poussé à mener une recherche spécialisée avec les contraintes traditionnelles de la RI et les contraintes d'un domaine en pleine évolution avec ses caractéristiques et ses lois : les réseaux d'entreprises. Les difficultés d'analyse du web afin de répondre au besoin du génie industriel (détecter les secteurs d'activités des entreprises et extraire leurs compétences) nous ont conduit à mettre en œuvre des techniques diverses : extraction d'information, fouille de données, ontologie, apprentissage par réseau de neurones, et à les intégrer au sein d'une architecture de traitement originale.

Partie 2 : Détection Automatique des Activités d'Entreprises

Dans cette partie, nous présentons un premier enjeu de la thèse, qui consiste à utiliser une ressource sémantique structurée propre au domaine. Cette ressource est un thésaurus inspiré du NAF (Nomenclature des Activités Françaises). Une approche basée sur des outils et des méthodes de recherche d'informations, à savoir une indexation contrôlée et une mesure de similarité, est étudiée. Elle est mise en place pour la détection automatique des secteurs d'activités des entreprises à partir de leur sites web. La bonne détection de l'activité d'une telle entreprise est l'une des clés permettant de faire émerger des réseaux coopératifs d'entreprises de divers types.

Cette partie est composée de trois chapitres. Le premier chapitre justifie le besoin de la recherche d'information pour la construction de réseaux d'entreprises. Ainsi il positionne notre problématique par rapport aux Organisations Virtuelles (OV) et VBE (Virtuel Breeding Environment). Le deuxième chapitre décrit l'approche de détection des secteurs d'activités des entreprises. Cette approche est composée par quatre étapes : Extraction, Lemmatisation, Indexation et Appariement. Une étude de ses performances est présentée. Dans un dernier chapitre, nous appliquons une méthode de construction des réseaux d'entreprises, basée sur un algorithme de clustering. Nous finissons par expliciter les limites de réseaux construits.

Chapitre 6

Problématique

6.1 Introduction

Le concept de l'organisation virtuelle (OV) représente un des exemples les plus discutés des réseaux de collaboration, qui a soulevé des espérances considérables dans beaucoup de domaines d'application (réseaux d'entreprises, les hôpitaux, les universités, les organisations gouvernementales etc). La possibilité de former rapidement une OV, déclenchée par une opportunité commerciale et spécifiquement conçue en fonction des conditions de cette occasion, est fréquemment mentionnée comme expression d'un mécanisme d'agilité et de survie face à la turbulence du marché. La même idée est également très attrayante dans d'autres contextes orientés affaires.
Dans la suite nous allons expliciter la problématique des organisations virtuelles dans le cadre de notre travail, montrant en particulier comment est justifié le besoin de la recherche et l'extraction d'information pour la construction des réseaux d'entreprises.

6.2 OV et VBE

Trouver le bon partenaire dans des conditions adéquates pour mettre en œuvre le processus de collaboration s'est avéré coûteux en termes de temps et effort. Notamment, les obstacles incluent le manque d'information (par exemple non-disponibilité des catalogues avec des profils normalisés avec les bonnes compétences et capacité), le manque d'infrastructure commune de collaboration et le manque de volonté des organismes de joindre le processus de collaboration. Tous ces obstacles ont poussé à chercher à mettre en œuvre une approche réaliste dans un cadre assisté par ordinateur pour aider à la création des organisations virtuelles. Trouver un partenaire, c'est trouver les bonnes conditions avec les bonnes informations (activité et compétence dans notre cas).
Si nous nous intèressons à cet objectif de recherche et d'extraction de l'information pertinente qui permettent la construction des réseaux, il y a en littérature beaucoup de recherche traitant les données caractéristiques sur les partenaires potentiels pour des organismes gérés en réseau [37] [131] [55]. D'autres approches [104] utilisent le site web des entreprises pour détecter des informations pertinentes : profile de l'entreprise, activité, adresse... Cependant, ces approches sont développées dans un environnement virtuel fermé (Virtual Breeding Environment). Ce VBE fournit déjà une présélection des partenaires potentiels, dans lesquels toutes les organisations donnent volontairement les données caractéristiques exigées. Au contraire,

l'approche que nous présentons dans ce mémoire est basée sur l'hypothèse d'un environnement ouvert des partenaires potentiels, de ce fait ayant une plus large application. En effet, le processus du choix de partenaires doit être basé sur l'utilisation d'information disponible publique non restreinte. Cette contrainte induit les mécanismes spécifiques d'extraction de l'information, que nous abordons ci-dessous.

6.3 Besoin de recherche et d'extraction d'information

Pour faciliter la coopération, ces organisations ont besoin d'une infrastructure leur permettant de partager des documents, de travailler et de communiquer ensemble malgré les contraintes géographiques. C'est pourquoi les organisations virtuelles s'appuient fortement sur les technologies de traitement de l'information.
Pour construire un système d'aide à la décision pour la gestion de la collaboration inter-organisations, les approches de recherche et d'extraction d'informations sont sollicitées pour mettre en exergue l'information caractérisant le réseau [37] [131] [55]. Ces approches de recherche et d'extraction d'information gèrent la création dynamique des organisations virtuelles. Comme nous l'avons vu en introduction, il existe deux types de recherche pour la gestion dynamique de ces organisations virtuelles :

- Une recherche dans un environnement fermé où les organisations se mettent d'accord d'avance pour travailler ensemble à court terme (pour une durée précise). Pour ce faire, elles partagent leurs connaissances et leurs informations (savoir faire, compétences...) sous un format donné et une structure homogène. Cette alliance est en général définie pour un court terme, une fois le bien ou le service est livré, le regroupement est dissocié. Ce type de réseau est caractérisé par des frontières très nettes dans lequel les nouveaux venus ne sont autorisés qu'en cas d'incident (Exemple : un partenaire quitte le réseau).
- Une deuxième recherche qui se fait dans un environnement ouvert où les organisations ne se connaissent pas et ont une information hétérogène publique et non restreinte. Ce type d'information rend la recherche plus difficile car on est face à des documents mal structurés caractérisés par un contenu hétérogène. Ce type de réseau est réalisé pour un nombre non prédéfini de processus, ce sont des alliances à caractère stratégique. Toutes les organisations intéressées et correspondantes aux objectifs du réseau peuvent y adhérer.

Notre travail se situe dans le deuxième type de recherche. Des travaux antérieurs au sein de notre équipe ont proposé une typologie des modes de coordination entre les différentes entreprises du réseau [130, 15]. Cette typologie est basée sur deux paramètres : la complémentarité des activités et la similarité des compétences. Ces deux paramètres ont été identifiés comme étant discriminants pour justifier le choix d'un type de coopération industrielle. C'est pourquoi notre besoin d'information s'articule autour de deux systèmes d'extraction d'information (complémentarité des activités et similarité des compétences). Nous nous limitons dans cette première

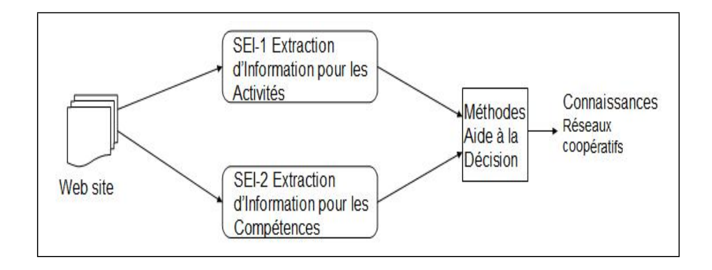


FIGURE 6.1 – Deux systèmes d'extraction d'information pour les entreprises

partie à la recherche et l'extraction d'information sur les activités et le savoir faire des entreprises. Nous proposons une approche basée sur des méthodes et outils de la recherche d'information. Les informations extraites sur les activités et les savoir faire, nous les utiliserons dans une deuxième étape pour montrer comment elles génèrent des nouvelles connaissances et permettent de faire émerger des propositions opérationnelles de mise en réseaux des entreprises.

6.4 Pourquoi le NAF

Connaitre l'activité principale d'une entreprise donnée est une question importante pour la gestion d'un réseau de collaboration. C'est aussi une question pertinente pour l'entreprise elle-même, pour savoir quels sont ses concurrents ou simplement pour s'assurer qu'elle met suffisamment d'information publique à propos de son activité, par exemple sur son site web.
Dans notre travail, nous utilisons un thésaurus qui reflète une représentation sémantique et conceptuelle de tous les domaines d'activités. Notre thesaurus est inspiré du NAF [1]. Il est utilisé en amont du moteur de recherche et sert de ressource sémantique externe pour améliorer l'expressivité du besoin d'information avant de le soumettre au système de recherche d'information. Cette technique peut s'avérer efficace, notamment lorsqu'il s'agit d'information traitant d'un domaine spécifique (activités des entreprises par exemple), dans la mesure où elle permet à l'utilisateur d'exprimer son besoin d'information dans un langage contrôlé. Nous effectuons la lemmatisation, avec l'outil TreeTagger [2], des termes du NAF ainsi qu'une élimination des mots vides. Le résultat est le vocabulaire contrôlé (VC) qui est un ensemble de termes.
Avec ces deux phases, nous voulons construire des vecteurs pour toutes les classes et les sous-classes du NAF, i.e C28, C28.1, C28.2, ...et construire un vecteur pour chaque site web d'entreprise. Pour cela on utilise les techniques traditionnelles de la RI, une représentation vectorielle des termes des libellés des classes et sous-classes NAF. Dans une troisième phase on effectue un appariement entre le vecteur classe et le vecteur entreprise pour mesurer le degré de rapprochement.

1. http ://www.insee.fr/fr/nom-def-met/nomenclatures/naf/pages/naf.pdf
2. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

6.5 Utilisation de l'information détectée sur les activités

La complémentarité des activités est l'un des deux paramètres qui a été identifié comme étant discriminant pour justifier le choix d'un mode de coordination industriel [15]. Le choix de ce paramètre est fondé sur l'analyse des activités pour extraire l'information pertinente qui permet d'établir un degré de rapprochement entre deux activités différentes. Ce degré de rapprochement entre les activités est utilisé ensuite comme paramètres de commande dans la création du réseau d'entreprises. Dans la suite nous allons explorer la notion de complémentarité entre les activités pour bien expliciter l'utilisation de l'information détectée sur les activités.

6.5.1 Définition de la complémentarité des activités dans un réseau d'entreprises

Avant de développer une analyse de la complémentarité des activités dans un réseau, il est nécessaire de préciser formellement ce que nous entendons par *complémentarité des activités*. Pour ce faire nous nous référons à des travaux d'économie industrielle et aux théories de l'organisation industrielle. Ainsi en théorie de l'organisation industrielle, Richardson [138] a défini la complémentarité de la manière suivante :
Deux activités sont qualifiées de complémentaires si elles correspondent à différentes phases successives d'un processus de production.
En économie industrielle la définition standard de la complémentarité s'inscrit dans une logique de marché [115] :
Des activités sont mutuellement complémentaires si l'augmentation de l'une de ces activités accroît la rentabilité marginale de toutes les autres activités du groupe. Cependant, nous déduisons que la définition de la complémentarité des activités est la suivante :
Complémentarité d'activités : quand les domaines d'activité de l'entreprise identifiés par des codes NAF interviennent plus ou moins fréquemment dans des produits intégrés (dans le secteur de la mécanique, produits intégrés : produits dont la conception et la réalisation requiert l'intervention conjointe de plusieurs domaines d'activité).
On pourra parler d'activités "supplémentaires" dans un secteur d'activités, quand il s'agit d'activités d'entreprises qui interviennent dans des produits généralement disjoints, qui n'offrent donc généralement pas d'opportunités de collaboration dans la réalisation de produit commun.

6.5.2 Modélisation de la complémentarité

Pour modéliser la complémentarité des activités nous utilisons la théorie des graphes qui offre l'avantage de faciliter la manipulation des objets et de leurs relations, avec une représentation graphique naturelle. L'ensemble des techniques et outils mathématiques mis au point en théorie des graphes permet de démontrer facilement des propriétés, d'en déduire des méthodes de résolution. En effet, la théorie des graphes offre un large panel de méthodes et algorithmes qui nous permettent

d'atteindre notre objectif. Elle permet aussi d'extraire des indicateurs représentatifs de la complémentarité des activités.
Dans les définitions données ci-dessus, la relation de complémentarité entre deux entreprises est symétrique. En effet, si une entreprise $E1$ est complémentaire d'une entreprise $E2$, cette dernière est forcément complémentaire de l'entreprise $E1$.
Dans notre travail, on constitue un graphe NAF qui modélise la complémentarité des activités où les éléments des graphes sont les secteurs NAFs, la liaison de complémentarité entre deux secteurs est représentée par un arc avec une évaluation. Ce graphe de complémentarité est construit manuellement par expertise [3].

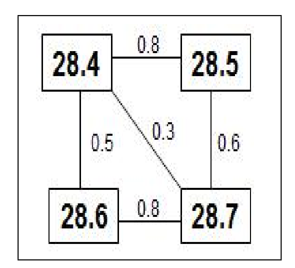


FIGURE 6.2 – Graphe de complémentarité entre les secteurs d'activités : 28.4 Forge, emboutissage, estampage ; métallurgie des poudres ; 28.5 Traitement des métaux ; mécanique générale ; 28.6 Fabrication de coutellerie, d'outillage et de quincaillerie ; 28.7 Fabrication d'autres ouvrages en métaux

L'intérêt du graphe NAF c'est qu'il est générique, et donc la complémentarité entre les champs d'activités peut être étudiée à partir d'expert du secteur d'activités, sans se reporter à une enquête spécifique à chaque entreprise. Dans des traveaux antérieurs de notre laboratoire, le recueil d'information sur la complémentarité supposait de connaître à l'avance les entreprises étudiées.

3. Xavier Boucher, Patrick Burlat (experts en génie industriel et modélisation des compétences d'entreprises) et le laboratoire de la mécanique de l'ENISE

Chapitre 7

Détection automatique des secteurs d'activités des entreprises

7.1 Introduction

Dans ce chapitre, nous cherchons à étudier la question des performances et de l'adéquation éventuelle des techniques de la recherche d'information dans une application spécifique à un domaine d'information métier ciblé (secteurs d'activités des entreprises). Le domaine métier est en premier lieu caractérisé par une complexité importante liée au fait que l'information s'y exprime de manière peu structurée : les textes composant le corpus ne suivent aucune structure standard ; la sémantique du vocabulaire utilisé est très lié au domaine métier (vocabulaire contextualisé) ; la structure linguistique des textes est parfois absente ; l'ensemble de ces facteurs induisent de forts risques d'ambiguïté. Mais le domaine métier est également caractérisé par un ensemble de spécificités dont on peut tirer parti de manière formelle, permettant de réduire cette complexité intrinsèque. Dans notre démarche de recherche, nous n'avons pas de réponse a priori sur l'efficacité des techniques de RI lorsqu'elles sont confrontées à la réalité de l'information métier : l'évaluation de leurs performances font partie de l'étude.

Dans le cadre des chapitres 7 et 8, le domaine d'information métier que nous ciblons peut être délimité par une double frontière. D'une part, il s'agit d'un secteur industriel particulier (l'industrie mécanique) tel que nous le précisons en section 6.2. D'autre part, il s'agit d'un type d'information spécifique : nous cherchons à identifier des informations caractérisant le domaine d'activité des entreprises. Ayant délimité ce "domaine informationnel métier", nous avons cherché à tirer parti de ses spécificités en cherchant des caractérisations de ce domaine, afin d'accroître l'efficience des dispositifs de RI : par quelle unité informationnelle est exprimée ce domaine (mot, expression ou phrase) ? Quelle granularité peut-on avoir sur les secteurs d'activités des entreprises ? Quelle ambigüité informationnelle et sémantique peut-on croiser dans ce domaine et par quelle ressource sémantique (taxinomie, thesaurus) peut-on guider la recherche ?

Le fait de cibler un domaine métier bien spécifique comme la mécanique nous permet de chercher des ressources sémantiques susceptibles de le caractériser. Nous avons sélectionné comme ressource sémantique le standard national Code NAF (Nomenclature des Activités Françaises), en limitant son utilisation au

domaine industriel de la mécanique. Le code NAF nous fournit une représentation conceptuelle hiérarchisée de tous les secteurs d'activités de ce domaine industriel : c'est une structure hiérarchique de classes et sous-classes de secteurs d'activités. Ce code NAF va être utilisé comme ressource sémantique externe, afin d'améliorer l'expressivité du besoin d'information avant de le soumettre au système de recherche d'information. L'intérêt du code NAF est qu'il délimite le domaine de recherche en explicitant ses caractéristiques et ses spécificités. Le système de détection des secteurs d'activités que nous réalisons traite des entreprises françaises, mais il est facilement exploitable à l'international pour tout pays francophone : la détection automatique du NAF permet de traiter toutes les entreprises, indépendamment du fait que leur NAF soit ou non répertorié dans les bases de données institutionnelles.

Dans notre recherche, le NAF est utilisé pour améliorer l'efficacité du processus d'indexation des sites web des entreprises. Il va servir à contrôler l'information qui circule dans le texte pour ne laisser passer que celle pertinente à notre domaine informationnel. Cette indexation conceptuelle tend à ne sélectionner que les plus importants concepts figurant dans le NAF, au contraire d'une indexation classique qui a pour but de couvrir tout le document. Parallèlement nous utilisons cet apport sémantique de manière plus large grâce aux techniques d'apprentissage par réseau de neurones en créant des liens sémantiques (synonymie, généralisation...) entre les termes du domaine.

Ce chapitre est structuré en 5 sections. La section 2 décrit les variables de l'environnement de recherche d'information à savoir le corpus et le code NAF. La section 3 décrit l'approche de détection des secteurs d'activités des entreprises. Nous mettrons l'accent sur l'usage du NAF en tant que ressource externe pour effectuer une indexation contrôlée et sur le processus d'appariement qui permet de mesurer la pertinence d'une classe NAF (secteur d'activité) vis-à-vis d'un site web d'une entreprise. Ainsi les sections 4 et 5 décrivent respectivement deux méthodes d'appariement qui ont été appliquées. Des mesures de performances de chacune de ces méthodes sont présentées suite à des tests d'évaluation.

7.2 Variables de recherche

7.2.1 Corpus d'expérimentation

Pour tester notre approche, nous avons sélectionné un ensemble de 100 entreprises. Ces entreprises sont axées sur le même secteur d'activité « la mécanique ». L'étude de ce secteur d'activité est motivée par la présence de plusieurs entreprises de ce secteur dans notre région, ainsi que la présence d'une base d'informations générales (Nom, URL, adresse...) sur ce type d'entreprises dans notre laboratoire. Il est ainsi possible d'utiliser des ressources sémantiques supplémentaire très ciblées. Ce secteur contient plusieurs sous-secteurs diversifiés. À partir des URLs de base de ces entreprises, l'ensemble des pages web est récupéré automatiquement au moyen de

l'aspirateur de site Web Wget [1]. Nous avons pu ainsi collecter 11926 pages HTML. Le corpus, qui est ainsi constitué des pages extraites des sites web présentant les entreprises, est extrêmement hétérogène et complexe (liens, images, texte mal structuré, animations, informations hétérogène, etc.). Cette information de départ (les sites web des entreprises), pertinente vis-à-vis de notre objectif d'extraction, se caractérise par une structuration instable avec un contenu informationnel très lié au domaine métier traité. Ce corpus pose des difficultés d'analyse qui conduisent à mettre en œuvre de nombreuses et diverses techniques informatiques.

7.2.2 Code NAF

C'est l'un des codes de l'INSEE [2]. Il permet la codification de l'activité principale exercée dans une entreprise ou une association. La NAF (nomenclature d'activités française) est une liste couvrant l'ensemble des activités économiques. On parlait de code APE (Activité Principale Exercée) de 1973 à 1992 et de code NAF depuis le 1er Janvier 1993. C'est un élément obligatoire sur un bulletin de salaire. Il est composé de 3 chiffres et une lettre. La figure 6.1 présente un extrait du code NAF que nous avons utilisé. Le code NAF est organisé en classes qui contiennent une ou plusieurs sous-classes. Chaque activité y est définie par un intitulé et repérée par un code, par exemple : « Fabrication de menuiseries et fermetures métalliques », dont le code NAF est 28.1C. Pour chaque entreprise ou établissement, l'INSEE détermine, en fonction des informations dont il dispose (résultats d'enquêtes ou déclarations de l'entreprise) et de règles de classement statistique, l'activité figurant dans la NAF qui correspond le mieux à son APE.
L'interêt du NAF est son statut de standard. L'impact de cet aspect est évident pour que les mécanismes d'extraction d'information soient susceptibles d'être beaucoup mieux acceptés et réutilisés. De plus l'information sémantique de base ainsi disponible fait déjà l'objet d'un certain consensus.
Il existe d'autres nomenclatures standard de classification des entreprises selon leurs activités. Par exemple le code KOMPASS est un code international qui propose des informations sur 2.1 millions entreprises et leurs produits dans 70 pays (nom, adresse, contacts téléphoniques, mail, le détail des produits et services proposés par l'entreprise). Le choix du NAF est justifié par les éléments suivants :

- Nous traitons dans cette thèse la langue française.
- D'un point de vue recherche, nous mettons au point des mécanismes qui peuvent tout à fait être transposés par la suite à d'autres ressources sémantiques de même types avec d'autres langues.
- La structure hiérarchique du NAF, qui se représente comme un arbre où dans chaque nœud on trouve une étiquette représentant le secteur ou le sous-secteur

1. http ://www.gnu.org/software/wget/wget.html
2. Un code Insee est un code numérique ou alphanumérique, élaboré par l'Institut national de la statistique et des études économiques, service public français chargé de la production et de l'analyse des différentes données statistiques concernant les collectivités, la géographie, les populations et les entreprises.

28.1C	Fabrication de menuiseries et fermetures métalliques
28.2	Fabrication de réservoirs métalliques et de chaudières pour le chauffage central
28.2C	Fabrication de réservoirs, citernes et conteneurs métalliques
28.2D	Fabrication de radiateurs et de chaudières pour le chauffage central
28.3	Chaudronnerie
28.3A	Fabrication de générateurs de vapeur
28.3B	Chaudronnerie nucléaire
28.3C	Chaudronnerie-tuyauterie
28.4	Forge, emboutissage, estampage ; métallurgie des poudres
28.4A	Forge, estampage, matriçage
28.4B	Découpage, emboutissage
28.4C	Métallurgie des poudres
28.5	Traitement des métaux ; mécanique générale
28.5A	Traitement et revêtement des métaux
28.5C	Décolletage
28.5D	Mécanique générale
28.6	Fabrication de coutellerie, d'outillage et de quincaillerie
28.6A	Fabrication de coutellerie
28.6C	Fabrication d'outillage à main
28.6D	Fabrication d'outillage mécanique
28.6F	Fabrication de serrures et de ferrures
28.7	Fabrication d'autres ouvrages en métaux
28.7A	Fabrication de fûts et emballages métalliques similaires
28.7C	Fabrication d'emballages métalliques légers
28.7E	Fabrication d'articles en fils métalliques
28.7G	Visserie et boulonnerie
28.7H	Fabrication de ressorts
28.7J	Fabrication de chaînes
28.7L	Fabrication d'articles métalliques ménagers
28.7N	Fabrication de petits articles métalliques
28.7Q	Fabrication d'articles métalliques divers
34	Industrie automobile
34.1	Construction de véhicules automobiles
34.1Z	Construction de véhicules automobiles
34.2	Fabrication de carrosseries et remorques
34.2	Carrosseries et remorques et services associés
34.20	Carrosseries et remorques et services associés
34.2A	Fabrication de carrosseries automobiles
34.2B	Fabrications de caravanes et véhicules de loisirs
34.3	Fabrication d'équipements automobiles
34.3	Equipements pour automobiles et services associés
34.3Z	Fabrication d'équipements automobiles
29.1A	Fabrication de moteurs et turbines
29.1B	Fabrication de pompes
29.1D	Fabrication de transmissions hydrauliques et pneumatiques
29.1E	Fabrication de compresseurs
29.1F	Fabrication d'articles de robinetterie
29.1H	Fabrication de roulements
29.1J	Fabrication d'organes mécaniques de transmission
29.2	Fabrication de machines d'usage général
29.2A	Fabrication de fours et brûleurs
29.2C	Fabrication d'ascenseurs, monte-charges et escaliers mécaniques
29.2D	Fabrication d'équipements de levage et de manutention
29.2F	Fabrication d'équipements aérauliques et frigorifiques industriels
29.2J	Fabrication d'appareils de pesage
29.2L	Fabrication de matériel pour les industries chimiques
29.2M	Fabrication d'autres machines d'usage général
29.3	Fabrication de machines agricoles
29.3A	Fabrication de tracteurs agricoles
29.3C	Réparation de matériel agricole
29.3D	Fabrication de matériel agricole
29.4	Fabrication de machines-outils
29.4A	Fabrication de machines-outils à métaux
29.4	Fabrication de machines-outils
29.4A	Fabrication de machines-outils à métaux
29.4C	Fabrication de machines-outils portatives à moteur incorporé
29.4B	Fabrication de machines-outils à bois
29.4D	Fabrication de matériel de soudage
29.4E	Fabrication d'autres machines-outils
29.5	Fabrication d'autres machines d'usage spécifique
29.5A	Fabrication de machines pour la métallurgie
29.5B	Fabrication de matériels de mines pour l'extraction
29.5D	Fabrication de matériels de travaux publics
29.6	Fabrication d'armes et de munitions
29.6A	Fabrication d'armement
29.6B	Fabrication d'armes de chasse, de tir et de défense
29.7	Fabrication d'appareils domestiques
29.7A	Fabrication d'appareils électroménagers
29.7C	Fabrication d'appareils ménagers non électriques

FIGURE 7.1 – Extrait du code NAF qu'on utilise (version 2003)

d'activité identifié par un code, rend facile l'exploitation et le traitement d'un point de vue informatique.

7.3 Approche de détection des secteurs d'activités

Compte tenu de l'existence d'une ressource sémantique très structurée comme le NAF, nous sommes conduits à utiliser des techniques d'indexation relativement classiques (indexation contrôlée) pour filtrer l'information qui circule dans le texte de l'entreprise. Rappelons toujours qu'en termes de recherche, la question consiste à montrer la valeur ajoutée de l'usage de ressources sémantiques propres au métier, par l'étude des performances finales du système. Pour l'extraction d'information sur les activités, nous avons procédé de manière statistique, en nous basant sur l'approche d'indexation contrôlée. Notre approche[76] [79] se déroule en trois phases décrites par la figure 7.2. Nous utilisons le thésaurus qui reflète une représentation sémantique et conceptuelle de tous les domaines d'activités. Dans notre cas le thesaurus est le code NAF 7.1.

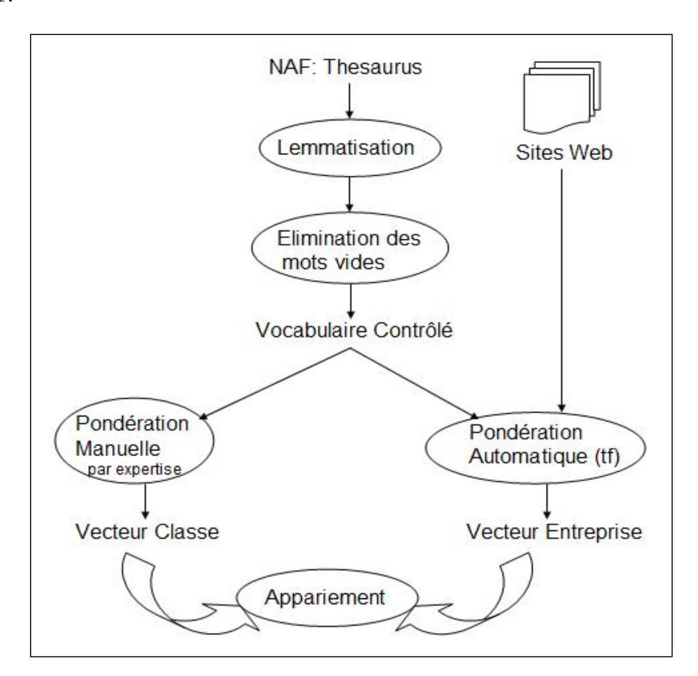


FIGURE 7.2 – Approche visée pour l'extraction des activités des entreprises

Notre thesaurus est utilisé en amont du moteur de recherche. Il sert de ressource sémantique externe pour améliorer l'expressivité du besoin d'information (quelle est mon code NAF à partir de mon site web ?) avant de le soumettre au système de recherche d'information. Cette technique peut s'avérer efficace, notamment lorsqu'il s'agit d'information traitant d'un domaine spécifique (activités des entreprises par exemple), dans la mesure où elle permet à l'utilisateur d'exprimer son besoin d'information dans un langage contrôlé. Nous effectuons la lemmatisation *(avec l'outil TreeTagger)* des termes du thesaurus ainsi qu'une élimination des mots vides. Le ré-

sultat est le Vocabulaire Contrôlé Hiérarchique (VCH) qui est un ensemble de termes (mots simples et mots composés), par exemple : usinage, emboutissage, machines-outils...
Dans une première phase une pondération manuelle est faite sur ce vocabulaire contrôlé ; elle permet d'attribuer, par expertise, un poids (1, 2 ou 3) pour chaque terme. Le poids d'un terme dans un document traduit l'importance de ce terme dans le document. En réorganisant l'ensemble des termes du VCH selon la structure initiale du NAF, nous obtenons un vecteur pour chaque classe NAF (vecteur classe).
Dans une deuxième phase, nous utilisons le VCH pour réaliser une pondération automatique du site web de l'entreprise. Cette pondération est basée sur le calcul de la fréquence du terme dans le texte du site de l'entreprise après avoir effectué un filtrage pour ne garder que les termes qui sont présents dans le VCH. Cette approche repose sur l'idée qu'il existe un rapport entre le contenu véhiculé par un texte et les mots utilisés dans le texte, que ce rapport est en fonction de la fréquence d'usage des mots, et qu'il existe une relation entre la capacité d'un mot à être choisi comme terme d'indexation et sa fréquence d'emploi.
Avec ces deux phases, nous voulons construire des vecteurs pour toutes les classes et les sous-classes du NAF, i.e C28, C28.1, C28.2, etc, et construire un vecteur pour chaque site web d'entreprise. Chaque vecteur est l'ensemble des descripteurs d'un document (classes ou sous-classes NAF) ou d'une requête (site web d'une entreprise) avec leurs pondérations (poids informationnels). Pour cela, on utilise les techniques traditionnelles de la RI et une représentation vectorielle des termes des libellés des classes et sous-classes NAF. Dans une troisième phase, on effectue un appariement entre le vecteur classe et le vecteur entreprise pour mesurer le degré de rapprochement.

7.3.1 Extraction

Au vu de nombreux exemples de pages web, dans lesquels l'information pertinente est noyée dans le texte dédié à la mise en forme ou à l'architecture du site web, nous avons vu naître le besoin d'établir des règles permettant d'extraire ce texte avec le moins de bruit possible. De façon analogue, des programmes spécifiques sont nécessaires pour extraire automatiquement de l'information dans les documents de type HTML, sans que celle-çi soit toujours explicitement structurée par un jeu de balises adéquates. C'est ce type de traitement que nous avons cherché à mettre en œuvre sur notre corpus. Nous avons utilisé le navigateur Lynx [3] qui est un programme de conversion de la version HTML en format texte. Ce programme [4] fonctionne par suppression et transformation de balises. Il prend en entrée un fichier ".html" ou ".htm" classique et propose en sortie la version en format ".txt". Nous

3. http ://lynx.browser.org/
4. Lynx est un navigateur "texte" trés connu dans le monde Unix (il existe aussi pour d'autres plateformes telles que Windows). On entend par navigateur texte, un navigateur qui affiche le contenu d'une page en mode texte, sans aucun rendu graphique. Il ne tient pas compte des feuilles de style, des balises de formatage (font...), des attributs de formatage et affiche tout avec une fonte unique, une taille unique de caractères.

avons en outre développé des programmes de nettoyage, qui normalisent le texte brut afin qu'il corresponde aux normes typographiques, et suppriment les éléments pouvant mettre en échec la suite du traitement.

7.3.2 Lemmatisation

L'analyse morphosyntaxique d'un discours de texte consiste à évaluer sa forme morphologique et la fonction grammaticale de ses éléments constitutifs. La morphologie est une branche de la linguistique qui étudie la façon dont les morphèmes (la plus petite unité porteuse de sens qu'il soit possible d'isoler dans un énoncé) se combinent pour former des lemmes (une unité autonome qui constitue la langue). Au cours de cette analyse morphosyntaxique, pour chaque mot on distingue sa catégorie grammaticale et son lemme. La lemmatisation désigne l'analyse lexicale du contenu d'un texte regroupant des mots d'une même famille. Chacun des mots se trouve réduit à une entité appelée lemme. La lemmatisation regroupe les différents formes que peut prendre un mot : le nom, le pluriel, le verbe à l'infinitif, etc.
Il existe plusieurs outils et plate-formes d'analyse morphosyntaxique. Celui que nous utilisons dans le cadre de notre travail est TreeTagger [5]. C'est un outil pour l'annotation grammaticale de données textuelles, qui associe à chacun des mots du discours son genre : noms, verbes, adjectifs, etc, et son lemme. Cet outil a été développé par Helmut Schmid dans le cadre du projet "TC" à l'institut de Linguistique informatique de l'Université de Stuttgart. TreeTagger a été utilisé avec succès pour différentes langues : allemand, anglais, français, italien, chinois. Il est fondé sur un algorithme d'arbre de décision pour effectuer l'analyse grammaticale.

7.3.3 Indexation

Cette étape est primordiale dans un processus de recherche d'informations. Elle consiste à analyser le document afin de produire un ensemble de mots clés, appelées aussi descripteurs, utilisés dans le processus de recherche d'informations. Nous avons effectué au début de notre expérience un premier test d'indexation, basé sur le calcul des fréquences des termes, dont le résultat était insatisfaisant vu que les termes clés qui représentent le document ne donnent pas l'information pertinente qui nous permet d'identifier l'activité de l'entreprise. En effet nous retrouvons beaucoup de termes qui ne sont pas pertinents pour notre recherche. C'est pourquoi dans un deuxième test nous avons effectué une indexation contrôlée par notre VCH, qui est faite en utilisant le NAF.

L'indexation contrôlée est composée de deux étapes : une première qui consiste en un filtrage pour ne conserver que les termes qui sont représentés par le VCH. Cette étape a pour but de maîtriser l'information qui circule dans le site web de l'entreprise et le cadrer par rapport au domaine traité. La deuxième étape consiste en une indexation traditionnelle basée sur la fréquence des termes qui sont

5. http ://perl.linguistes.free.fr/telechargements.html

filtrés. Cette dernière est faite à l'aide du logiciel d'indexation SMART (System for the Mechanical Analysis and Retrieval of Text) appelé aussi Salton's Magic Automatic Retrieval Technique, qui est un système d'indexation pour la recherche d'informations.

Depuis les années 1970, des chercheurs se sont penchés sur l'intérêt d'utiliser des ressources lexico-sémantiques dans le processus d'indexation. L'intérêt se justifie par le souci d'un meilleur contrôle et une uniformalisation du langage d'indexation. Ces resources ont été utilisées avec succès pour améliorer les performances des systèmes de recherche d'informations dans différentes applications [81] [93] [71]. L'utilisation du VCH pour réaliser une indexation contrôlée a pour objectif de pénaliser les termes porteurs d'ambiguïté qui ont un impact direct sur la performance du système. La deuxième raison de l'utilisation du VCH est l'exploitation de la force informationnelle et représentative que constitue le NAF, comme un référentiel standard du domaine, pour explorer le contenu des sites web des entreprises.

7.3.4 Appariement

L'objectif des systèmes de recherche d'informations (SRI) est d'établir une correspondance entre l'information recherchée par un utilisateur et celle contenue dans leur base documentaire. Pour y parvenir, ces systèmes font un appariement des termes de la requête posée avec ceux représentant le contenu des documents. Dans notre approche, les requêtes et les documents sont représentés dans l'espace vectoriel engendré par les termes d'indexation [147] en utilisant le système SMART.
Dans notre cas, l'appariement (mesure de similarité) consiste à retrouver les vecteurs documents (les classes et les sous-classes du NAF) qui s'approchent le plus du vecteur requête (vecteur entreprise). La phase d'appariement se déroule en deux étapes : dans un premier temps on cherche à détecter la classe du NAF la plus pertinente pour l'entreprise ; dans un deuxième temps on explore les sous-classes de cette classe pour détecter de nouveau une sous-classe. Ce processus d'appariement document-requête permet de mesurer la pertinence d'un document vis-à-vis d'une requête.
Plusieurs techniques classiques de la RI sont disponibles pour répondre à ce besoin d'appariement. Avant toute remise en cause de ces techniques, notre problématique de recherche vise dans un premier temps à vérifier si l'usage de ressources sémantiques propres au métier permet d'enrichir suffisamment la performance issue de ce type de techniques. En nous appuyant sur la mise à disposition du code NAF, nous avons donc décidé de tester plusieurs de ces mécanismes d'appariement. Dans un premier temps, des appariements basés sur un modèle vectoriel (mesure de produit scalaire, cosinus et mesure de Jaccard) et dans un deuxième temps, des appariements basés sur un modèle connexionniste en mettant en place un réseau de neurones. Ces deux méthodes d'appariement seront développés respectivement dans les sections qui se suivent.

7.4 Mesure de similarité simple

Dès le départ, une fonction de similarité de type td-idf nous a semblé mal adaptée dans notre cas. Car les documents de la collection ont des petites tailles. Notre choix de fonctions de similarité (appariement) s'est porté sur les trois fonctions principales de la RI :
Le produit scalaire :

$$RSV(Q, D_j) = \sum_{i=1}^{N} q_i \times d_{ij}$$

La mesure Cosinus :

$$RSV(Q, D_j) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{(\sum_{i=1}^{N} q_i^2)^{(1/2)} \times (\sum_{i=1}^{N} d_{ij}^2)^{(1/2)}}$$

La mesure de Jaccard :

$$RSV(Q, D_j) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{\sum_{i=1}^{N} q_i^2 + \sum_{i=1}^{N} d_{ij}^2 - \sum_{i=1}^{N} q_i \times d_{ij}}$$

Pour pouvoir expérimenter notre modèle de recherche, nous avons extrait une sous-collection de notre corpus de test. Cette sous-collection est composée de 25 entreprises « requêtes », de 20 classes et de sous-classes NAF « documents »et d'une base de données créée manuellement listant les documents censés être pertinents pour chaque requête. Les documents sont classés principalement en trois classes C28 (Travail des métaux), C29 (Fabrication de machines et d'équipements) et C34 (industrie automobile). Dans la suite, nous allons présenter les résultats correspondant aux différents modèles de similarité.

7.4.1 Mesure avec le produit scalaire

Req\Doc	NAF Réel	C28	C28.1	C28.2	C28.3	C28.4	C28.5	C28.6	C28.	C29	C29.1	C29.2	C29.3	C29.4	C29.5	C29.6	C29.7	C34	C34.1	C34.2	C34.3	NAF Déte	Pertinence
E1	C29.1	47	4	12	8	4	15	12	6	79	55	14	4	2	14	0	19	21	22	5	6	C29.1	P
E2	C34.3	143	3	18	66	45	0	30	17	20	27	6	6	6	9	6	0	83	29	45	46	C28.3	NP
E3	C29.1	103	4	13	21	14	49	34	14	25	23	2	2	14	8	2	0	59	45	18	33	C28.5	NP
E4	C34.3	12	1	3	5	7	0	3	3	9	2	2	2	4	10	2	0	13	10	6	2	C34.1	SP
E5	C34.3	58	2	12	28	16	3	19	11	25	17	12	2	6	4	2	10	40	15	33	22	C28.3	NP
E6	C28.5	16	1	3	3	1	6	8	3	14	10	6	2	4	10	2	0	15	6	9	9	C28.6	SP
E7	C74.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C74.2	P
E8	C28.5	120	0	15	22	35	73	18	15	4	5	2	2	0	2	0	0	16	14	15	17	C28.5	P
E9	C28.4	28	2	3	8	10	10	6	7	5	4	2	2	2	4	2	0	14	5	5	13	C28.4	P
E10	C28.4	896	33	221	222	322	139	224	156	375	51	12	0	52	118	0	288	464	355	149	225	C28.4	P
E11	C28.4	336	13	12	23	11	279	19	74	41	9	2	2	2	36	2	0	51	33	36	4	C28.5	SP
E12	C28.5	82	5	12	11	4	61	16	10	18	19	21	16	10	16	10	0	22	5	19	6	C28.5	P
E13	C28.5	100	7	10	9	7	72	21	13	35	38	30	30	14	36	14	0	33	4	29	2	C28.5	P
E14	C28.4	18	0	0	3	13	9	0	0	12	6	0	0	0	6	0	3	7	6	2	6	C28.4	P
E15	C28.5	21	0	0	3	16	9	0	0	12	6	0	0	0	6	0	3	7	6	2	6	C28.4	SP
E16	C28.5	40	6	10	12	2	20	15	10	16	22	14	14	12	17	12	0	37	8	30	12	C28.5	P
E17	C28.5	75	31	23	23	20	28	7	33	13	4	4	4	2	14	2	0	24	29	6	5	C28.5	P
E18	C28.5	52	3	1	1	9	32	16	6	39	32	22	17	4	17	0	8	21	16	6	10	C28.5	P
E19	C28.5	98	2	2	2	15	89	37	5	15	13	12	10	6	13	4	2	10	1	8	3	C28.5	P
E20	C29.2	50	1	3	9	6	33	14	2	15	2	2	2	2	16	2	0	26	29	11	0	C28.5	NP
E21	C34.1	173	3	98	62	6	0	7	5	239	288	2	2	4	2	2	9	222	162	61	166	C29.1	NP
E22	C34.2	54	0	25	26	1	0	7	9	129	92	0	54	0	37	0	0	151	108	101	3	C34.1	SP
E23	C34.3	50	11	15	13	10	8	9	12	14	14	4	7	6	6	4	0	22	7	11	12	C28.2	NP
E24	C34.1	21	1	10	4	3	5	4	6	35	7	3	21	0	18	6	3	52	43	11	3	C34.1	P
E25	C34.3	106	34	15	20	36	32	13	23	17	15	8	11	12	13	8	0	51	16	35	5	C28.4	NP

FIGURE 7.3 – Résultats de mesure de similarité obtenus avec la fonction produit scalaire

La première colonne représente l'entreprise (la requête) avec son code NAF réel que nous cherchons à retrouver (le document pertinent). Le code NAF réel est composé d'une classe NAF et une sous-classe, en conséquence deux informations pertinentes que le système est censé retrouver. La première ligne représente notre collection de documents (20 documents constitués de 3 classes et 17 sous-classes). La matrice des valeurs représente les degrés de rapprochement (scores) entre le vecteur de la requête (l'entreprise) et le vecteur document (classe ou sous-classe).
Dans chaque ligne, le score le plus élevé représente le degré maximum de rapprochement entre les deux vecteurs. Si nous prenons la pemière entreprise avec son code NAF réel le C29.1, le score le plus élevé (score maximal) parmi les trois classes (C28, C29, et C34) est détecté sur le document C29 (c'est la bonne classe). De même le score le plus élevé des sous-classes de la classe C29 est le C29.1 (c'est la bonne sous-classe). En conclusion pour cette première entreprise le système a retrouvé le bon code NAF (le code NAF détecté est égal au code NAF réel). A chaque requête on évalue la pertinence de la réponse par les trois niveaux :

- P : Une réponse Pertinente c'est-à-dire qu'on a détecté la bonne classe et la bonne sous-classe NAF.
- SP : Une réponse Semi-Pertinente c'est-à-dire qu'on a détecté la bonne classe NAF mais pas la sous-classe.
- NP : Une réponse Non Pertinente c'est-à-dire qu'on n'a pas détecté ni la bonne classe ni la bonne sous-classe NAF.

7.4.2 Mesure avec la fonction cosinus

Req\Doc	NAF Réel	C28	C28.1	C28.2	C28.3	C28.4	C28.5	C28.6	C28.7	C29	C29.1	C29.2	C29.3	C29.4	C29.5	C29.6	C29.7	C34	C34.1	C34.2	C34.3	NAF Détecté	Pertinence
E1	C29.1	0,16	0,03	0,1	0,05	0,03	0,12	0,07	0,02	0,3	0,4	0,09	0,05	0,02	0,09	0	0,22	0,11	0,12	0,03	0,05	C29.1	P
E2	C34.3	0,14	0	0,04	0,13	0,09	0	0,05	0,02	0,02	0,05	0,01	0,02	0,02	0,01	0,01	0	0,17	0,05	0,11	0,14	C34.3	P
E3	C29.1	0,23	0,02	0,07	0,1	0,07	0,27	0,14	0,04	0,06	0,11	0	0,1	0,12	0,03	0,01	0	0,22	0,17	0,09	0,21	C28.5	NP
E4	C34.3	0,12	0,02	0,07	0,1	0,16	0	0,05	0,04	0,1	0,04	0,04	0,07	0,15	0,19	0,06	0	0,2	0,16	0,14	0,05	C34.1	P
E5	C34.3	0,14	0,01	0,07	0,14	0,09	0,01	0,08	0,03	0,06	0,08	0,05	0,01	0,05	0,01	0,01	0,08	0,16	0,05	0,18	0,14	C34.2	SP
E6	C28.5	0,17	0,02	0,07	0,06	0,02	0,16	0,16	0,04	0,17	0,23	0,13	0,15	0,16	0,18	0,07	0,07	0,26	0,1	0,22	0,26	C34.3	NP
E7	C74.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C74.2	P
E8	C28.5	0,18	0	0,05	0,07	0,12	0,27	0,05	0,03	0	0,01	0	0,01	0	0	0	0	0,04	0,03	0,05	0,07	C28.5	P
E9	C28.4	0,14	0,02	0,03	0,08	0,12	0,12	0,05	0,04	0,02	0,04	0,02	0,03	0,03	0,03	0,03	0	0,11	0,04	0,05	0,18	C28.4	P
E10	C28.4	0,19	0,01	0,11	0,1	0,16	0,07	0,09	0,04	0,09	0,02	0	0	0,04	0,04	0	0,21	0,16	0,12	0,07	0,13	C34.3	NP
E11	C28.4	0,25	0,02	0,02	0,03	0,01	0,51	0,02	0,07	0,03	0,01	0	0	0	0,05	0	0	0,06	0,04	0,06	0	C28.5	SP
E12	C28.5	0,25	0,03	0,09	0,07	0,02	0,46	0,09	0,04	0,06	0,12	0,13	0,18	0,11	0,09	0,09	0	0,11	0,02	0,13	0,05	C28.5	P
E13	C28.5	0,18	0,03	0,04	0,03	0,03	0,32	0,07	0,03	0,07	0,14	0,11	0,2	0,09	0,12	0,08	0	0,09	0,01	0,12	0,01	C28.5	P
E14	C28.4	0,13	0	0	0,04	0,22	0,16	0	0	0,09	0,09	0	0	0	0,08	0	0,07	0,08	0,07	0,03	0,12	C28.4	P
E15	C28.5	0,15	0	0	0,04	0,27	0,16	0	0	0,09	0,09	0	0	0	0,08	0	0,07	0,08	0,07	0,03	0,12	C28.4	P
E16	C28.5	0,18	0,06	0,11	0,11	0,02	0,22	0,12	0,06	0,08	0,21	0,12	0,23	0,2	0,14	0,17	0	0,27	0,05	0,31	0,15	C34.2	NP
E17	C28.5	0,19	0,2	0,14	0,12	0,12	0,18	0,03	0,11	0,03	0,02	0,02	0,03	0,01	0,06	0,01	0	0,1	0,12	0,03	0,03	C28.1	SP
E18	C28.5	0,21	0,03	0	0	0,08	0,31	0,11	0,03	0,2	0,27	0,17	0,25	0,06	0,12	0	0,1	0,13	0,1	0,05	0,11	C28.5	P
E19	C28.5	0,16	0	0	0	0,05	0,36	0,11	0,01	0,02	0,04	0,04	0,06	0,03	0,04	0,02	0,01	0,02	0	0,03	0,01	C28.5	P
E20	C29.2	0,13	0	0,01	0,05	0,03	0,22	0,07	0	0,04	0,01	0,01	0,01	0,02	0,08	0,01	0	0,11	0,12	0,06	0	C28.5	NP
E21	C34.1	0,08	0	0,11	0,06	0	0	0	0	0,12	0,29	0	0	0	0	0	0,01	0,17	0,12	0,06	0,22	C34.3	SP
E22	C34.2	0,04	0	0,05	0,05	0	0	0,01	0,01	0,13	0,18	0	0,18	0	0,06	0	0	0,22	0,16	0,21	0	C34.2	P
E23	C34.3	0,29	0,16	0,21	0,16	0,14	0,11	0,09	0,09	0,09	0,17	0,04	0,15	0,13	0,06	0,07	0	0,21	0,06	0,15	0,19	C28.2	NP
E24	C34.1	0,1	0,01	0,11	0,04	0,03	0,06	0,03	0,04	0,19	0,07	0,02	0,37	0	0,16	0,09	0,05	0,42	0,34	0,12	0,04	C34.1	P
E25	C34.3	0,27	0,21	0,09	0,1	0,21	0,2	0,06	0,08	0,04	0,08	0,04	0,1	0,11	0,06	0,06	0	0,21	0,06	0,2	0,03	C28.1	NP

FIGURE 7.4 – Résultats de mesure de similarité obtenus avec la fonction cosinus

L'entreprise E7 est un test pour la robustesse du système, c'est une entreprise qui n'appartient pas aux trois classes NAF choisies et sa similarité est effectivement nulle avec toutes les classes et sous-classes.

7.4.3 Mesure avec la fonction Jaccard

Req\Doc	NAF Réel	C28	C28.1	C28.2	C28.3	C28.4	C28.5	C28.6	C28.7	C29	C29.1	C29.2	C29.3	C29.4	C29.5	C29.6	C29.7	C34	C34.1	C34.2	C34.3	NAF Détecté	Pertinence
E1	C29.1	0,08	0,01	0,04	0,02	0,01	0,05	0,03	0,01	**0,17**	**0,22**	0,04	0,01	0	0,04	0	0,08	0,06	0,06	0,01	0,02	C29.1	P
E2	C34.3	**0,04**	0	0	**0,02**	0,01	0	0,01	0	0	0	0	0	0	0	0	0	0,03	0,01	0,01	0,01	C28.3	NP
E3	C29.1	**0,13**	0	0,02	0,03	0,02	0,1	**0,06**	0,02	0,03	0,04	0	0	0,02	0,01	0	0	0,1	0,07	0,03	0,06	C28.6	NP
E4	C34.3	0,02	0,01	0,03	0,04	0,07	0	0,02	0,01	0,02	0,01	0,01	0,03	0,08	0,07	0,03	0	**0,07**	**0,06**	**0,06**	0,02	C34.1	SP
E5	C34.3	**0,07**	0	0,02	**0,05**	0,03	0	0,03	0,01	0,03	0,03	0,02	0	0,01	0	0	0,02	0,06	0,02	0,07	0,04	C28.3	NP
E6	C28.5	0,03	0,01	0,03	0,02	0,01	0,07	0,06	0,01	0,04	0,01	0,05	0,08	0,08	0,07	0,03	0,03	**0,09**	0,03	0,1	**0,13**	C34.3	NP
E7	C74.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C74.2	P
E8	C28.5	**0,09**	0	0,01	0,02	0,03	**0,07**	0,01	0,01	0	0	0	0	0	0	0	0	0,01	0,01	0,01	0,01	C28.5	P
E9	C28.4	0,05	0,01	0,01	0,04	0,06	0,06	0,02	0,02	0,01	0,02	0,01	0,01	0,01	0,01	0,01	0	**0,06**	0,02	0,03	**0,09**	C34.3	NP
E10	C28.4	**0,01**	0	0	**0,02**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C28.3	SP
E11	C28.4	**0,07**	0	0	0	0	**0,06**	0	0,01	0	0	0	0	0	0	0	0	0,01	0	0	0	C28.5	SP
E12	C28.5	**0,14**	0,01	0,03	0,03	0,01	**0,23**	0,04	0,02	0,03	0,05	0,06	0,05	0,03	0,04	0,03	0	0,05	0,01	0,06	0,01	C28.5	P
E13	C28.5	**0,09**	0	0,01	0,01	0	**0,09**	0,02	0,01	0,03	0,04	0,03	0,04	0,01	0,04	0,01	0	0,03	0	0,03	0	C28.5	P
E14	C28.4	**0,4**	0	0	0,02	**0,12**	0,08	0	0	0,03	0,04	0	0	0	0,03	0	0,03	0,03	0,03	0,01	0,06	C28.4	P
E15	C28.5	**0,04**	0	0	**0,2**	0,15	0,08	0	0	0,03	0,04	0	0	0	0,03	0	0,03	0,03	0,03	0,01	0,06	C28.3	SP
E16	C28.5	0,08	0,03	0,05	0,06	0,01	0,11	0,06	0,03	0,03	0,11	0,06	0,1	0,08	0,07	0,08	0	**0,15**	0,03	**0,18**	0,07	C34.2	NP
E17	C28.5	**0,1**	**0,07**	0,05	0,05	0,04	**0,07**	0,01	0,06	0,01	0	0	0,01	0	0,03	0	0	0,04	0,06	0,01	0,01	C28.5	P
E18	C28.5	**0,1**	0,01	0	0	0,04	0,17	0,06	0,01	**0,1**	0,15	0,09	0,1	0,02	0,06	0	0,04	0,07	0,05	0,02	0,05	C28 C29	NP
E19	C28.5	**0,08**	0	0	0	0,01	**0,1**	0,03	0	0,01	0,01	0,01	0,01	0	0,01	0	0	0	0	0	0	C28.5	P
E20	C29.2	**0,07**	0	0	0,02	0,01	**0,08**	0,03	0	0,02	0	0	0	0	0,03	0	0	0,05	0,06	0,02	0	C28.5	NP
E21	C34.1	0,01	0	0	0	0	0	0	0	**0,02**	0,02	0	0	0	0	0	0	**0,02**	0,01	0	0,01	C29 C34	NP
E22	C34.2	0,01	0	0	0	0	0	0	0	0,04	0,03	0	0,01	0	0,01	0	0	**0,05**	0,02	**0,03**	0	C34.2	P
E23	C34.3	**0,11**	0,09	0,12	0,08	0,07	0,06	0,05	0,04	0,03	0,09	0,02	0,07	0,06	0,03	0,03	0	**0,11**	0,03	0,08	0,1	C28 C34	NP
E24	C34.1	0,04	0	0,06	0,02	0,01	0,03	0,01	0,01	0,08	0,03	0,01	0,18	0	0,09	0,04	0,02	**0,26**	**0,2**	0,06	0,01	C34.1	P
E25	C34.3	**0,15**	**0,08**	0,03	0,04	0,07	0,07	0,02	0,04	0,02	0,03	0,01	0,02	0,03	0,02	0,01	0	0,1	0,03	0,08	0,01	C28.1	NP

Figure 7.5 – Résultats de mesure de similarité obtenus avec la fonction Jaccard

Comme nous pouvons le constater, les valeurs fournies par la fonction Jaccard sont trés faibles, ce qui rend difficile la sélection des documents pertinents

7.4.4 Evaluation : analyse critique

L'évaluation de la performance du système est basée sur le calcul des deux indicateurs Précision et Rappel. Nous avons testé les trois fonctions traditionnelles, mais nous n'avons retenu que les résultats de la fonction cosinus (plus performante que les fonctions produit scalaire ou Jaccard suite à nos expérimentations).

Pour chaque vecteur entreprise, il existe seulement deux vecteurs classes pertinents (la classe et la sous-classe NAF). Notre objectif est d'augmenter la

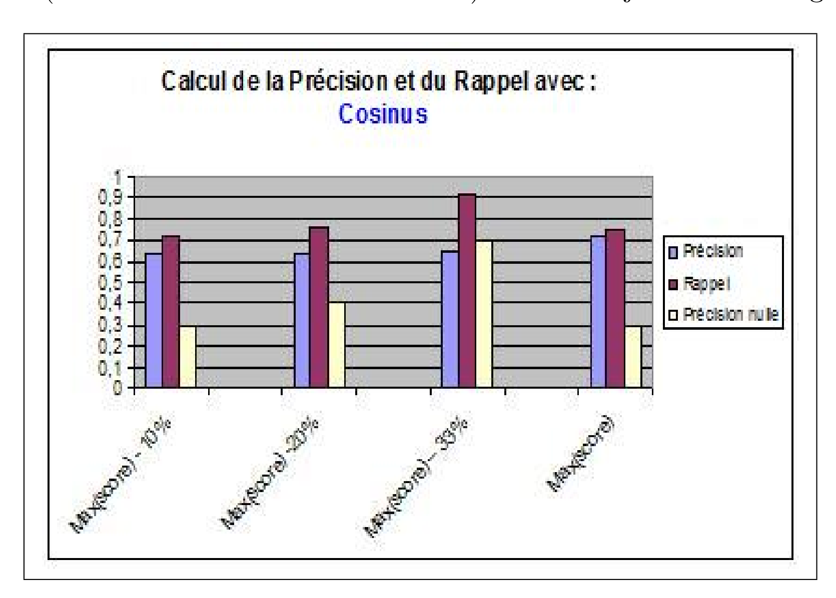


FIGURE 7.6 – Evaluation de la fonction Cosinus

précision du système ainsi que son rappel, mais aussi d'éviter le plus possible d'avoir des valeurs de précision nulles qui signifient que le système ne retrouve pas de documents pertinents. Notre étude diffère des cas traditionnels du domaine de la RI : par exemple nous avons un nombre de requêtes supérieur au nombre de documents. En effet du point de vue du Génie Industriel, la question est plutôt de savoir, pour une entreprise donnée, quel est son code NAF et non pas de prendre un code NAF et de chercher toutes les entreprises qui le possèdent.

Comme décrit dans le chapitre 2, la précision est l'intersection de l'ensemble des documents pertinents avec les documents retrouvés. Le rappel est l'intersection de l'ensemble des documents pertinents avec les documents retrouvée par rapport aux documents pertinents. L'ensemble des classes retournées est pris en compte suivant 3 intervalles qui dépendent du score maximal : [scoremax -α% score max, scoremax] avec $\alpha = 10$, 20 et 33. Les résultats montrent une bonne performance pour $\alpha = 33$. Sur cet intervalle, nous obtenons une performance de 0,5 de précision, 0,91 de rappel et 0,7 de précision non nulle. Ceci peut s'expliquer par le fait que nous

attribuons trois valeurs de poids aux termes des documents NAF lors de la phase de pondération (section 3.1). Cette technique d'évaluation est inspirée de la technique dite "évaluation à n documents prêts" qui a l'avantage de restituer l'ensemble des documents retrouvés par le système.

7.5 Mesure de similarité par réseau de neurones

Le modèle connexionniste par réseaux de neurones est une alternative au modèle vectoriel. L'approche connexionniste permet d'apporter plusieurs fonctionnalités souhaitées dans la recherche d'informations. Elle va d'un simple appariement des requêtes et des documents à des techniques associatives de documents pour l'expansion de la réponse (sélection de nouveaux documents). Dans notre problématique, ce modèle est particulièrement intéressant car il va permettre une représentation enrichie des ressources sémantiques propres au métier en vue d'augmenter leur valeur ajoutée dans le système d'extraction.

7.5.1 Définition des Réseaux de Neurones

Comme décrit dans la partie état de l'art (chapitre 2 section 3) les réseaux de neurones ont contribué à un modèle de recherche d'informations connue sous le nom « modèle connexionniste »[26] [119]. Les réseaux de neurones formels sont des structures, simulées par des algorithmes, qui tirent leur inspiration du fonctionnement élémentaire du systèmes nerveux. Ils sont très utilisés pour le traitement de l'information pour des applications de modélisation de la langue [16] [167], la reconnaissance de la parole [150], la recherche d'information [26] [119] [92], etc. La théorie des réseaux des neurones est issue de l'observation du fonctionnement du réseau de neurones biologiques et constitue un domaine de recherche en pleine effervescence.

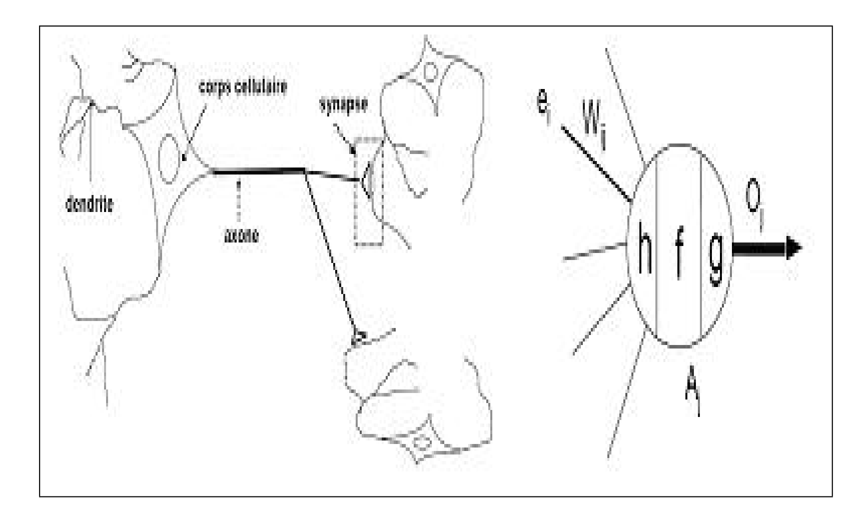


FIGURE 7.7 – représentation d'un neurone formel [111]

e_i : entrée du neurone ; A_j : activation du neurone ; O_j : sortie du neurone ; W_{ij} :

poids (synaptiques) ; h : fonction d'entrée ; f : fonction d'activation (ou transfert) ; g : fonction de sortie

La première modélisation d'un neurone a été présentée par Mac Culloch et Pitts en 1943 [111]. Ils ont proposé le modèle suivant : « le neurone formel fait une somme pondérée des potentiels d'activation $e_1, e_2...e_n$ qui lui parviennent, puis s'active suivant la valeur de cette somme pondérée. Si cette somme dépasse un certain seuil, le neurone est activé et transmet une réponse, si le neurone n'est pas activé il ne transmet rien »[56]. D'une façon générale le neurone formel [111] [63] est un processeur qui applique une opération simple à ses entrées et que l'on peut relier à d'autres, pour former un réseau qui peut réaliser une relation entrée-sortie quelconque, et d'une façon usuelle il calcule une somme pondérée et applique à cette somme une fonction de transfert non linéaire (échelon, sigmoïde, gaussienne, etc.)
Un réseau de neurones est un ensemble de neurones formels interconnectés et évoluant dans le temps par interactions réciproques. Son fonctionnement se base sur le comportement de chaque neurone (fonction d'activation) et l'interaction entre neurones (la structure et le poids des connexions). Une fois l'architecture et la dynamique du réseau choisis, le réseau va subir à son entrée les exemples à apprendre (phase d'apprentissage), l'algorithme d'apprentissage détermine la façon d'ajuster les poids du réseau pour obtenir la sortie désirée pour un exemple donné. La phase suivante est appelée phase d'utilisation ou de test. Elle consiste à présenter des exemples autres que ceux qui ont contribué à son apprentissage (des exemples généralement bruités ou incomplets). Pendant cette phase, le réseau va réagir selon les connaissances acquises durant la phase d'apprentissage.

7.5.2 Techniques d'apprentissage

Un des aspects importants caractérisant les réseaux de neurones est leur capacité à apprendre. L'apprentissage va permettre au réseau de modifier sa structure interne (poids synaptiques) pour s'adapter à son environnement [122] [148]. A chaque choix de coefficients synaptiques (poids de connexions) correspond alors un système, et c'est dans l'ensemble de ces systèmes que l'on se propose de trouver celui résolvant au mieux le problème. Pour pouvoir évaluer un système particulier, nous effectuons une série d'expériences permettant à chaque fois d'observer le comportement du réseau. Une expérience consiste à présenter un exemple d'entrée au système et la réponse est fournie à la sortie du système. L'évaluation du réseau se fait à chaque fois en examinant la valeur de la fonction d'erreur. Le processus d'apprentissage consiste alors à trouver un réseau minimisant cette fonction d'erreur[165]. L'apprentissage consiste, à partir d'exemples ou de prototypes fournis au réseau, à modifier les connexions à travers leurs poids de telle sorte que pour ces exemples, le réseau réponde correctement. Le pouvoir de généralisation du réseau de neurones lui permet alors de répondre même dans des cas non appris.

Les procédures d'apprentissage peuvent se subdiviser elles aussi en deux grandes

catégories : apprentissage supervisé et apprentissage non supervisé :
Apprentissage supervisé : Ce processus implique l'existence d'un « professeur »qui peut évaluer le succès ou l'échec du réseau quand il lui est présenté un stimulus (exemple) connu. On dit ainsi que ce stimulus fait partie de la base d'apprentissage. Cette supervision permet de renvoyer au réseau une information pour faire évoluer ses poids ou ses connexions afin de diminuer le taux d'échec. Cette information est une mesure de l'erreur commise exemple par exemple. La difficulté majeure de l'apprentissage est d'identifier les étapes du processus qui sont responsables de l'échec ou du succès (credit assignement problem).

Apprentissage non supervisé : Il s'agit de donner au réseau une quantité suffisante d'exemples contenant des corrélations pour que le réseau règle ses poids automatiquement. Cette architecture correspond bien à une forme de supervision. En effet, à travers ce type d'apprentissage, on cherche à imposer au système un fonctionnement spécifique à partir des données. Le réseau commence à apprendre en modifiant ces poids synaptiques. L'adaptation s'effectue à partir d'un algorithme d'optimisation. L'initialisation des poids est le plus souvent aléatoire.

Dans notre réseau de neurones, l'apprentissage est supervisé, car il nécessite une intervention pour évaluer la réponse pertinente. Cette intervention consiste à fournir des couples à apprendre (entrée, sortie désirée) dont les valeurs de sortie sont pondérées selon leurs pertinences vis-à-vis l'entrée. A partir du vecteur souhaité [6] fourni par l'utilisateur, le réseau calcule pour chaque exemple la fonction coût et réinjecte l'erreur. Ainsi, les poids synaptiques seront modifiés. Si l'erreur commise par le réseau est inférieure à un certain seuil, il est stable.

7.5.3 Présentation de l'architecture du réseau

Il n'existe pas de méthodes automatiques pour choisir l'architecture du réseau. Elle varie en fonction de l'application et dépend fortement des données à utiliser pour l'apprentissage. L'architecture du réseau est construite en fonction du nombre de couches à utiliser et du nombre de neurones dans chaque couche. Les neurones peuvent être organisés de différentes manières, ce qui définit l'architecture et le modèle du réseau [165].

Le réseau de neurones que nous utilisons [77] est le Multi Layer Perceptron qui est organisé en couches où l'information circule dans un seul sens. Le choix de ce type de réseau est justifié par le fait que nous avons trois types de données (requête, VCH, documents). Ceux-ci doivent être représentés séparément dans le réseau avec une logique d'emplacement. La requête est la clé du besoin d'information. Elle doit

6. Vecteur qui évalue la pertinence de chaque document en fonction de la réponse pertinente attendue par l'utilisateur. Aprés la première simulation, nous gardons les sorties des 18 neurones documents et nous augmentons les valeurs des deux neurones documents pertinents (classes et sous-classes)

être en entrée du réseau. La réponse est fonction des documents. Donc ces derniers doivent être à la sortie. Entre les deux se place le VCH qui contrôle l'association des termes entre la requête et les documents. Dans notre cas, il est constitué de trois couches (couche d'entrée, couche cachée et couche de sortie) avec deux types de neurones (neurone terme et neurone document). Les neurones termes correspondent à l'entrée du réseau. Car c'est en fonction d'eux que la requête est exprimée. Les neurones documents correspondent à la sortie du réseau pour exprimer la réponse en fonction du document le plus pertinent. Le processus suit un mécanisme de propagation d'activation. Autrement dit, un vecteur entreprise (requête) active initialement certaines cellules termes. Cette activation se propage vers les documents à travers les connexions entre les couches. Enfin, la connaissance peut évoluer par apprentissage. Dans le but de tester l'apprentissage en utilisant la statistique de pondération des termes et la statistique combinée avec la sémantique (relations de synonymie, de généralisation...), nous avons établi deux modèles : un modèle de base et un modèle enrichi.

7.5.3.1 Modèle de base

Le modèle de base illustré dans la figure 7.8 est constitué d'abord par une couche d'entrée qui représente une couche virtuelle liée à l'entrée du système et ne contient aucun neurone. Elle est créée dynamiquement à chaque interrogation (nouvelle requête). La couche suivante est constituée par n neurones de termes (n : nombre de termes du VC). Il existe un lien synaptique reliant chaque terme de la couche d'entrée (requête) à un terme de la couche des termes reflétant le poids de ce terme dans la requête. La dernière couche est celle de sortie constituée par m neurones de documents (m : nombre de documents, m classes et sous-classes NAF). Les scores des neurones documents sont directement les sorties du système. Le réseau reçoit à son

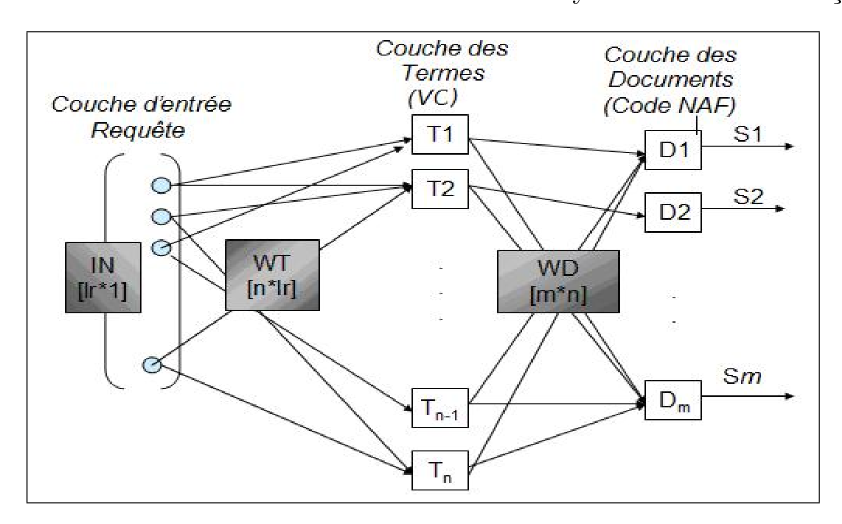


FIGURE 7.8 – Organisation des couches dans notre modèle de base

entrée un vecteur de termes (vecteur requête) activant ainsi les termes de la requête sur la couche des termes. Ensuite ces termes activés vont propager leurs activations

à leurs voisins. Enfin les termes activés directement à partir de la couche d'entrée et ceux activés par propagation vont envoyer leurs signaux d'activation vers la couche de sortie. Les documents sur la couche de sorties recevant des signaux pour être activés se déclenchent pour construire la réponse à la requête d'entrée. Les relations suivantes expliquent le processus d'activation :

$$\forall t \in T/t \in \overrightarrow{q}, E_i^T(\tau = 0) = freq_i$$

$$\forall t \in T/t \notin \overrightarrow{q}, E_i^T(\tau = 0) = 0$$

$$\forall t \in D, E_i^D(\tau = 0) = 0$$

$$E_i^D(\tau = 1) = \sum_{i \in T} E_t^T(\tau = 0).q_t.w_{d,t}$$

où
T : est la couche des neurones termes.
D : est la couche des neurones documents.
$freq_t$: est la fréquence absolue du terme t dans la requête q.
q_t : est le poids du terme t dans la requête q.
$w_{d,t}$: est le poids du terme t dans le document d.
Les deux premières équations représentent l'état initial des neurones termes (à $\tau = 0$). La troisième équation représente l'état initial des neurones document. L'état des documents (à $\tau = 1$) est représenté par la dernière équation : c'est la somme des produits de l'importance des termes activés, calculée à partir de leur fréquence absolue et leur fréquence relative.
Une fois l'architecture du réseau de neurones choisie, il est nécessaire d'effectuer un apprentissage. L'apprentissage détermine les valeurs des poids permettant à la sortie de réseau de neurones d'être aussi proche que possible de la réponse pertinente attendue (pour chaque entreprise détecter la classe et la sous-classe pertinentes). Cet apprentissage s'effectue en calculant à chaque fois l'écart de l'erreur entre le vecteur sortie du réseau et le vecteur désiré qui contient les scores qui privilégient la classe et la sous-classe pertinentes. L'erreur est rétropropagée à chaque fois dans les couches du réseau jusqu'à obtenir le résultat désiré, c'est-à-dire obtenir un réseau stable.

7.5.3.2 Modèle Enrichi

L'objectif du modèle enrichi est de tirer un meilleur parti des ressources sémantiques propres au métier. Le VCH, constitué précédemment, ne va plus être utilisé comme une ressource linguistique brute permettant de filtrer l'information. Mais il va être analysé pour faire apparaitre des relations d'ordre linguistique concernant la synonymie, la généralisation, la co-occurrence. L'un des intérêts de ce modèle enrichi est qu'il permet la représentation et l'usage de ces informations à forte valeur ajoutée.

Nous avons mis l'accent sur l'importance de la fonction de mise en correspondance et l'indexation qui ont un rôle majeur dans la performance du processus pour limiter le silence et le bruit de notre système, et ce pour nous garantir la sélection des documents les plus pertinents. La détection des termes en commun entre un document et une requête quelconque n'est pas satisfaisante par rapport à notre souci de ressortir tous les documents pertinents. Il nous faut donc aller un peu plus loin et ajouter des mécanismes complémentaires.
Notre objectif est toujours d'améliorer la fonction de mise en correspondance parce qu'un terme peut apparaître dans plusieurs documents et peut représenter plus qu'un concept. Par conséquent, si un document et une requête n'ont pas la même représentation, ce dernier ne sera pas retourné ce qui accroît le silence.

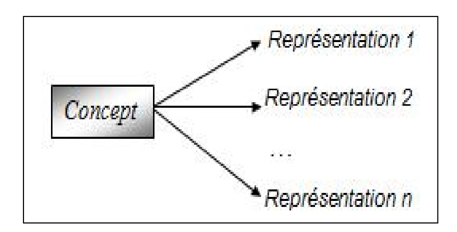


FIGURE 7.9 – Multiple représentations d'un concept par différents termes

En outre, un document peut être indexé par des termes spécifiques et éventuellement par des termes génériques. Cette problématique ne peut pas être résolue par une simple comparaison des représentations. Pour mettre en œuvre ces mécanismes, des relations entre les termes sont nécessaires.
Le modèle enrichi est une extension du modèle de base. Ce modèle est basé sur l'utilisation de la combinaison sémantique des termes [26] [119]. Nous avons inclus des relations de synonymie, de généralisation et de Co-occurrence. Chaque relation est représentée par une couche cachée. Tous les termes sont les mêmes dans toutes les couches. En passant d'une couche à la seconde, un nombre plus important de termes est activés. Notre modèle utilise des différents liens entre les termes basés

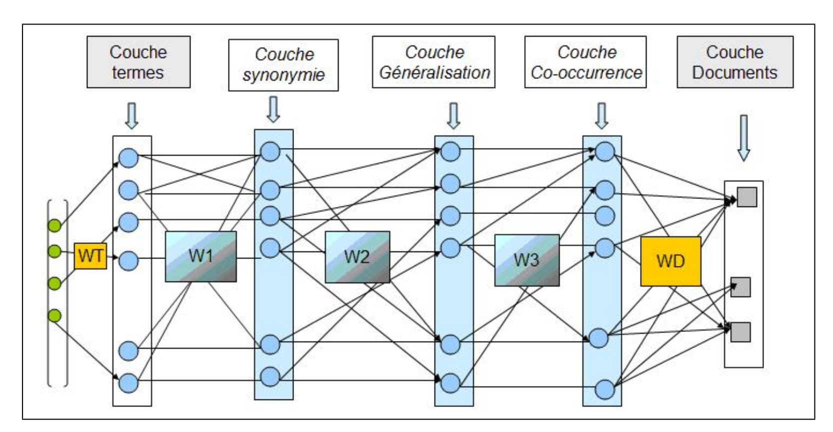


FIGURE 7.10 – Organisation des couches dans le modèle Enrichi

sur des relations statistiques et sémantiques. Les relations sémantiques sont expri-

mées essentiellement par des liens de synonymie et de généralisation. Les relations statistiques sont exprimées par des liens de co-occurrence.
Deux termes t_1 et t_2 sont synonymes s'ils représentent les mêmes concepts. Inversement, les relations sont symétriques, réflexives et transitives. Un terme t_1 généralise un autre terme t_2 si tous les concepts de t_2 peuvent être représentés par le terme t_1. Nous pouvons dire alors que le terme t_2 est spécifique du terme t_1. Cette relation est non symétrique et paradoxale. Deux termes t_1 et t_2 sont co-occurrents s'ils apparaissent ensemble en liaison forte au moins dans l'indexation d'un document. On associe à ces liens la fréquence d'apparition de ces termes dans tout le corpus c'est-à-dire dans les 20 documents (classes et sous-classes NAF).
Un neurone calcule son statut selon l'état de neurones qui sont reliés et selon les connexions impliquées. Le tableau 7.1, nous présentons un extrait descriptif des couches utilisées.

Synonymie		Généralisation	
Voiture	véhicule	mécanique	fraise
Chaudière	chaudronnerie	mécanique	décolletage
Réservoir	citerne	revêtement	métaux
Fabrication	construction	usinage	découpage

Table 7.1 – Exemple des termes de la couche synonymie et généralisation

La couche de co-occurrence est obtenue automatiquement en calculant le nombre d'apparitions des deux termes ensemble dans les documents. La décision, si deux termes sont synonymes, est liée au langage du domaine traité. La sémantique des termes dans ce domaine peut être différente de la sémantique des termes dans la langue française.

7.5.4 Performance du modèle connexionniste

En utilisant la formule de propagation proposée dans le modèle connexionniste, nous obtenons presque les mêmes documents restitués avec le même degré de pertinence que le modèle vectoriel.

7.6 Synthèse : comparaison du modèle vectoriel et connexionniste

Les résultats ci-dessus mettent en évidence que la précision est légèrement meilleure pour la fonction cosinus, mais que le modèle connexionniste a permis d'améliorer le rappel et l'indicateur de précision nulle. Concernant la comparaison entre ces 2 appariements, d'autres expérimentations seraient nécessaires dans le futur pour obtenir des conclusions plus définitives. En revanche, nous pouvons d'ores et déjà confirmer que la capacité à identifier correctement un code NAF est améliorée. Ainsi, ces techniques de recherche d'informations s'avèrent efficaces lorsqu'elles

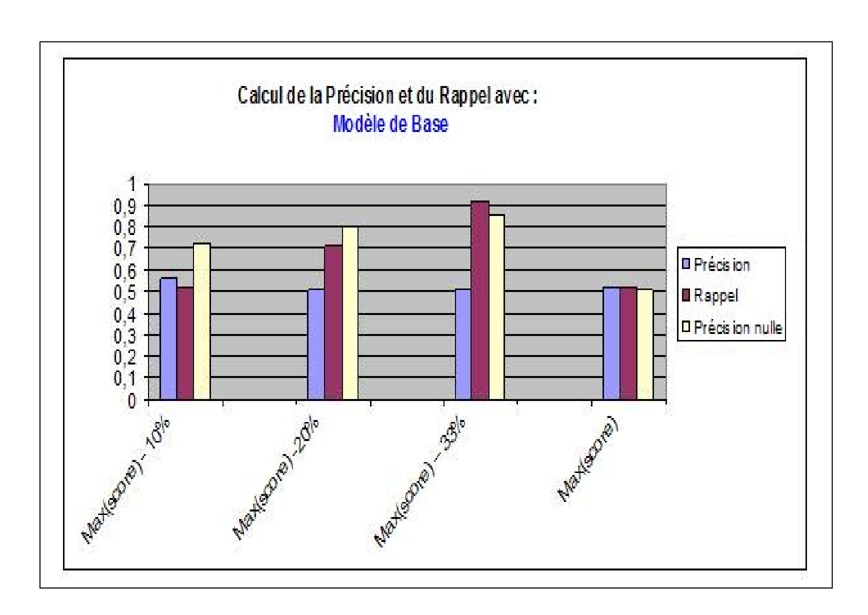


FIGURE 7.11 – Evaluation du modèle de base

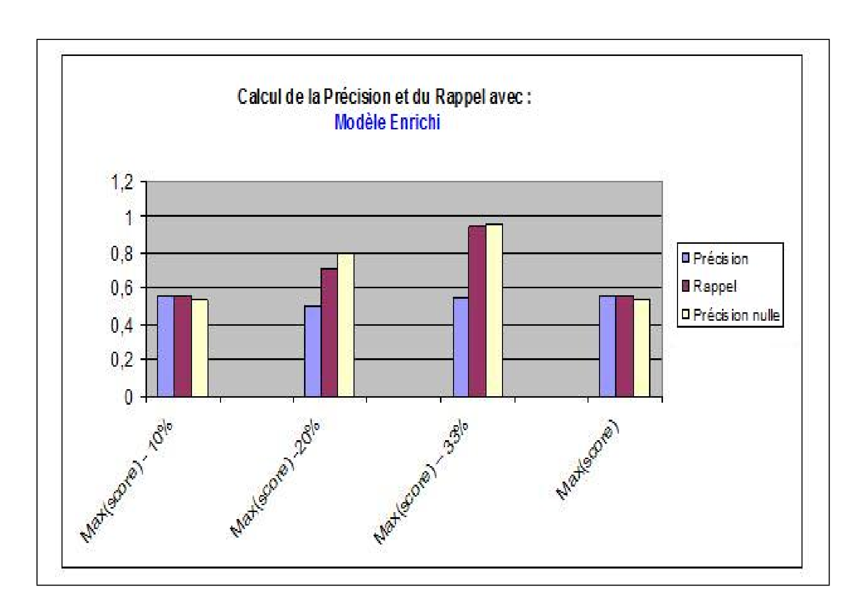


FIGURE 7.12 – Evaluation du modèle enrichi

sont enrichies par l'usage d'une ressource sémantique externe spécifique au métier, du type du code NAF.

	Modèle Vectoriel (Cosinus)	**Modèle connexionniste de base**	**Modèle connexionniste Enrichi**
Précision	0.64	0.51	0.55
Rappel	0.91	0.92	0.95
Précision nulle	0.3	0.14	0.04
Pourcentage de bonne réponse pour les classes	92%	80%	88%
Pourcentage de bonne réponse pour les sous classes	76%	72%	88%

Table 7.2 – Tableau récapitulatif d'évaluation des résultats

Le modèle connexionniste enrichi a permis d'améliorer la précision nulle : le nombre de cas pour lesquels on arrive à détecter la bonne classe NAF pour l'entreprise. De façon générale, pour tous les modèles, la précision moyenne est acceptable sans être très bonne. Ceci peut s'expliquer par le fait que les documents (surtout les sous-classes NAF d'une même classe) sont très proches les uns des autres. Une similarité document/document a été établie et elle montre bien cette proximité (figure 7.13). Ce qui rend difficile pour tout modèle la détection exacte de la bonne sous-classe NAF. C'est d'ailleurs la raison pour laquelle nous nous arrêtons au deuxième niveau du code NAF (classes et sous-classes directes).

Dans le modèle connexionniste enrichi, le fait de rajouter une couche de synonymie a renforcé la proximité entre les documents (sous-classes du NAF). Ce qui engendre une détérioration légère de la précision par rapport au modèle vectoriel.
Sur la base de ces résultats, nous constatons que les performances du modèle enrichi se rapprochent des performances du modèle vectoriel (calculé avec la fonction cosinus). A ce jour, nous ne pouvons pas juger réellement la performance du modèle connexionniste vis-à-vis du modèle vectoriel. Il faut mener une étude plus approfondie. Nous signalons aussi que le modèle connexionniste présente l'avantage sur la possibilité de l'apprentissage dynamique d'offrir un apprentissage à court terme (mécanisme de réinjection de la pertinence) qui permet une amélioration des résultats et un apprentissage à long terme du fait que le modèle est capable d'apprendre parfaitement un ensemble de requêtes.
En général, les travaux de recherche dans ce domaine montrent bien une bonne performance du modèle connexionniste vis-à-vis du modèle vectoriel [26] [119] [97]. Toutefois l'évaluation du système reste dépendante de la base des documents (taille des documents et nombre des documents). Dans notre cas les documents sont courts (de 5 à 50 termes) ce qui explique la faible existence (ou même l'absence) des rela-

Doc\Doc	C28	C28.1	C28.2	C28.3	C28.4	C28.5	C28.6	C28.7	C29	C29.1	C29.2	C29.3	C29.4	C29.5	C29.6	C29.7	C34	C34.	C34.2	C34.3
C28	1	0,29	0,49	0,31	0,43	0,42	0,43	0,35	0,05	0,02	0,01	0,09	0,01	0,05	0,01	0,03	0,06	0	0,11	0,02
C28.1	0,29	1	0,1	0,09	0	0,23	0,01	0,13	0	0,02	0,02	0,04	0,04	0,02	0,04	0	0,06	0	0,04	0
C28.2	0,49	0,1	1	0,33	0	0,05	0,08	0,1	0	0,02	0,02	0,04	0,04	0,02	0,03	0	0,08	0	0,2	0,03
C28.3	0,31	0,09	0,33	1	0,01	0,05	0,05	0,08	0	0,02	0,02	0,03	0,03	0,01	0,03	0	0,06	0	0,09	0,01
C28.4	0,43	0	0	0,01	1	0	0	0	0,03	0	0	0	0	0,05	0	0	0	0	0,01	0
C28.5	0,42	0,23	0,05	0,05	0	1	0,03	0,1	0	0	0	0	0	0	0	0	0	0	0	0
C28.6	0,43	0,01	0,08	0,05	0	0,03	1	0,06	0,03	0,03	0,03	0,16	0,03	0,03	0,02	0	0,07	0	0,1	0,03
C28.7	0,35	0,13	0,1	0,08	0	0,1	0,06	1	0,08	0,04	0,04	0,02	0,07	0,01	0,05	0,08	0,03	0	0,07	0,01
C29	0,05	0	0	0	0,03	0	0,03	0,08	1	0,47	0,49	0,22	0,24	0,52	0,33	0,3	0,09	0,1	0,03	0,09
C29.1	0,02	0,02	0,02	0,02	0	0	0,03	0,04	0,47	1	0,08	0,15	0,15	0,07	0,06	0	0,13	0,1	0,1	0,21
C29.2	0,1	0,02	0,02	0,02	0	0	0,03	0,04	0,49	0,08	1	0,14	0,07	0,07	0,06	0,06	0,04	0	0,06	0
C29.3	0,09	0,04	0,04	0,03	0	0	0,16	0,02	0,22	0,15	0,14	1	0,13	0,23	0,11	0	0,22	0,1	0,12	0
C29.4	0,01	0,04	0,04	0,03	0	0	0,03	0,07	0,24	0,15	0,07	0,13	1	0,13	0,11	0	0,15	0,1	0,12	0,15
C29.5	0,05	0,02	0,02	0,01	0,05	0	0,03	0,01	0,52	0,07	0,07	0,23	0,13	1	0,05	0	0,12	0,1	0,07	0
C29.6	0,01	0,04	0,03	0,03	0	0	0,02	0,05	0,33	0,06	0,06	0,11	0,11	0,05	1	0	0,07	0	0,11	0
C29.7	0,03	0	0	0	0	0	0	0,08	0,3	0	0,06	0	0	0	0	1	0	0	0	0
C34	0,06	0,06	0,08	0,06	0	0	0,07	0,03	0,09	0,13	0,04	0,22	0,15	0,12	0,07	0	1	0,7	0,51	0,31
C34.1	0,02	0,03	0,02	0,01	0	0	0,02	0,01	0,08	0,05	0	0,13	0,09	0,09	0	0	0,72	1	0,03	0,23
C34.2	0,11	0,04	0,2	0,09	0,01	0	0,1	0,07	0,03	0,1	0,06	0,12	0,12	0,07	0,11	0	0,51	0	1	0,09
C34.3	0,02	0	0,03	0,01	0	0	0,03	0,01	0,09	0,21	0	0	0,15	0	0	0	0,31	0,2	0,09	1

FIGURE 7.13 – Similarité Document/Document avec la fonction Cosinus

tions d'associations entre les termes des documents et les termes des VCH.

7.7 Conclusion

Nous avons présenté une contribution de détection automatique des activités d'entreprises. Cette contribution (SEI-1) présente un système automatique d'extraction d'information sur les activités d'entreprises à partir de leurs sites web. Elle est basée sur des méthodes et des outils de recherche d'information. Les mesures de similarité utilisées s'appuient sur les indicateurs standards de la RI (Précision et Rappel) et montrent une performance autour de 80% de bonnes réponses. Cependant la complémentarité des activités d'entreprises est insuffisante pour regrouper correctement les entreprises d'un même réseau de coopération. C'est pourquoi nous avons besoin du second système d'extraction concernant cette fois les compétences des entreprises (SEI-2).

Dans le chapitre suivant nous allons discuter les performances des outils utilisés dans ce premier système. Ainsi nous alllons explorer les limites de la construction d'un réseau d'entreprises en tenant compte uniquement de la complémentarité d'activités.

Chapitre 8

Application aux réseaux d'entreprises

8.1 Discussion sur les performances des outils utilisés

Dans cette première partie des contributions de la thèse, nous nous situons dans un contexte où l'information est mal structurée et bruitée. Cependant, nous avons pu l'exploiter grace à une ressources sémantique externe bien structurée et à forte valeur ajoutée. Cela répond à la problématique générale qui vise à faire évoluer l'enrichissement des techniques classiques de la recherche et de l'extraction d'information en utilisant des ressources propres au domaine métier. En même temps, on répond à un besoin d'information dans un contexte d'une application d'aide à la décision pour les collaborations inter-entreprises. Techniquement, nous avons démontré qu'il existe des solutions techniques en l'occurrence d'une indexation contrôlée et de réseaux sémantiques enrichis qui permettent de tirer partie efficacement des ressources sémantiques propres au métier.

Nous avons fait le choix de faire évaluer les possibilités de tirer parti de cette information métier au sein de deux techniques de la recherche d'information : une technique d'indexation basée sur le modèle vectoriel et d'autres techniques basées sur l'apprentissage par des modèles connexionnistes. Le but de ce choix était de sélectionner des méthodes présentant des atouts bien distincts, offrant un traitement statistique où l'importance du terme est déterminée en fonction de sa fréquence d'utilisation dans le texte, et un traitement sémantique basé sur la sémantique des termes pour exploiter la sémantique des textes dans la représentation de l'information. Cette deuxième technique est utilisée pour rendre possible une extension de la représentation des documents (ou requêtes) via les différentes relations sémantiques qu'elle implique.

Les performances globales du système ont montré une réussite de détection des secteurs d'activités des entreprises, autour de 92% sur les classes et de 76% sur les sous-classes pour la collection des entreprises testée par l'usage de ces techniques assez classiques. Ce résultat amène à conclure que si une ressource sémantique externe bien structurée est disponible, il ne semble pas utile d'avoir recours à des techniques de recherche d'information plus complexes tel qu'un apprentissage par réseau de neurones. Cependant l'avantage de l'un ou de l'autre en termes de performance n'est pas établi dans notre cas. Cela supposerait sans

doute une mise au point des modèles connexionnistes plus poussée et un protocole d'expérimentation élargi (voir à prendre en compte tout le NAF) sans oublier de souligner un premier avantage intéressant des modèles connexionnistes qu'est l'apprentissage dynamique.

Dans ce chapitre nous allons appliquer les résultats trouvés par le premier mécanisme d'extraction d'information qui s'occupe de la détection des secteurs d'activités des entreprises pour la construction des réseaux d'entreprises. Dans la suite, nous allons appliquer la méthode formelle de construction de réseaux, discuter ses résultats, et expliciter les limites des réseaux construits sur le seul résultat des informations pertinentes extraites par ce premier mécanisme.

8.2 Discussion sur l'application aux réseaux d'entreprises

Le travail décrit dans les sections suivantes est limité à un test de faisabilité. Il s'agit de montrer, par un exemple concret, que les informations issues du mécanisme d'extraction d'information (SEI-1) discuté auparavant sont effectivement utilisables dans un but d'aide à la décision pour la construction de réseaux d'entreprises.

8.2.1 Génération d'un graphe de complémentarité

L'objectif final de ce premier système est la détection potentielle des opportunités de collaboration entre différentes entreprises. A ce stade cette détection est basée seulement sur la complémentarité des activités, qui est un facteur important dans cette collaboration [115]. Nous nous référons à une méthode d'aide à la décision définie dans [31]. Les auteurs proposent une aide à la prise de décision basée sur les algorithmes de groupement qui peuvent être appliqués en utilisant les résultats de nos mécanismes d'extraction comme entrée.
L'information extraite par le SEI-1 est un secteur d'activité identifié par un code NAF. Pour appliquer la méthode d'aide à la décision, l'information utile est une évaluation de la complémentarité entre les secteurs d'activités, exprimée par un degré de complémentarité. Compte tenu du caractère générique du code NAF, nous avons proposé de compléter l'arborescence hiérarchique par une matrice générique d'indices de complémentarité entre les secteurs d'activités. Pour construire cette matrice de complémentarité des secteurs d'activités, nous avons eu recours à un recueil d'expertise auprès du domaine métier.

Notre modèle utilise les notions suivantes :
- L'activité, qui décrit une vue externe de l'entreprise. Nous décrivons ce que fait l'entreprise, ce qui aboutit aux produits et aux services qu'elle offre sur le marché. En quelque sorte l'activité s'exprime sous forme d'actions (on peut exprimer une activité par un verbe) ou par les produits et les services générés

par l'entreprise..
- La complémentarité d'activités : quand les domaines d'activités de deux entreprises identifiés par des codes NAF interviennent plus ou moins conjointement dans la réalisation d'un tel produit.
- La relation de complémentarité d'activité est symétrique.
- Le poids des arcs liant deux secteurs d'activités est compris entre 0 et 1.

L'intérêt du graphe NAF est qu'il est générique. La complémentarité entre les secteurs d'activités peut donc être étudiée à partir des connaissances d'experts du secteur d'activités, sans se reporter à une enquête spécifique à chaque entreprise (dans les travaux de [31], le recueil d'informations sur la complémentarité supposait de connaître à l'avance les entreprises étudiées). L'expert évalue à chaque fois si deux domaines seront complémentaires et à quelle degré de complémentarité (poids de l'arc). Nous avons demandé à 2 experts du domaine d'évaluer les degrés de complémentarités pour la partie du code NAF concernée. Cela nous a permis de converger vers la matrice de complémentarité présentée dans le tableau 8.1. Ce recueil est suffisant pour le test de faisabilité ciblé [1].

A ce stade la difficulté est de positionner chaque entreprise dans son domaine d'activité. C'est là que nous utilisons les résultats de notre système de détection des activités des entreprises. Si nous confrontons les résultats de notre système de détection automatique des activités sur nos 25 entreprises avec le graphe de complémentarité des activités (GCA). nous obtenons le graphe suivant (figure 8.1 : Nos 25 entreprises sont distribuées sur 8 secteurs d'activités. Une entreprise représentative est choisie pour chaque secteur.

- E1 : C29.1 Fabrication d' équipements mécaniques
- E2 : C28.4 Forge. emboutissage. estampage ; métallurgie des poudres
- E3 : C28.5 Traitement des métaux ; mécanique générale
- E4 : C34.1 Construction de véhicules automobiles
- E5 : C34.2 Fabrication de carrosseries et remorques
- E6 : C34.3 Fabrication d'équipements automobiles
- E7 : C74.3 Activités de contrôle et analyses techniques
- E8 : C29.2 Fabrication de machines d'usage général

8.2.1.1 Utilisation d'un algorithme de *clustering*

Nous réutilisons l'algorithme proposé par [31]. L'objectif de cet algorithme est d'isoler des sous-graphes fortement interconnectés en minimisant la perte d'information (perte d'arcs, perte de complémentarité potentielle). Les sous-graphes obtenus à la fin du processus de partitionnement représenteront les entreprises très complémentaires qui permettront de justifier d'une relation de type "réseau proactif" ou de type "firme". L'algorithme est basé sur un partitionnement, et il prend en compte plusieurs aspects spécifiques du graphe de complémentarité des

1. l'application d'un protocole plus systématique de recueil d'expertise ne poserait aucun problème, mais simplement sort du cadre utile à la thèse

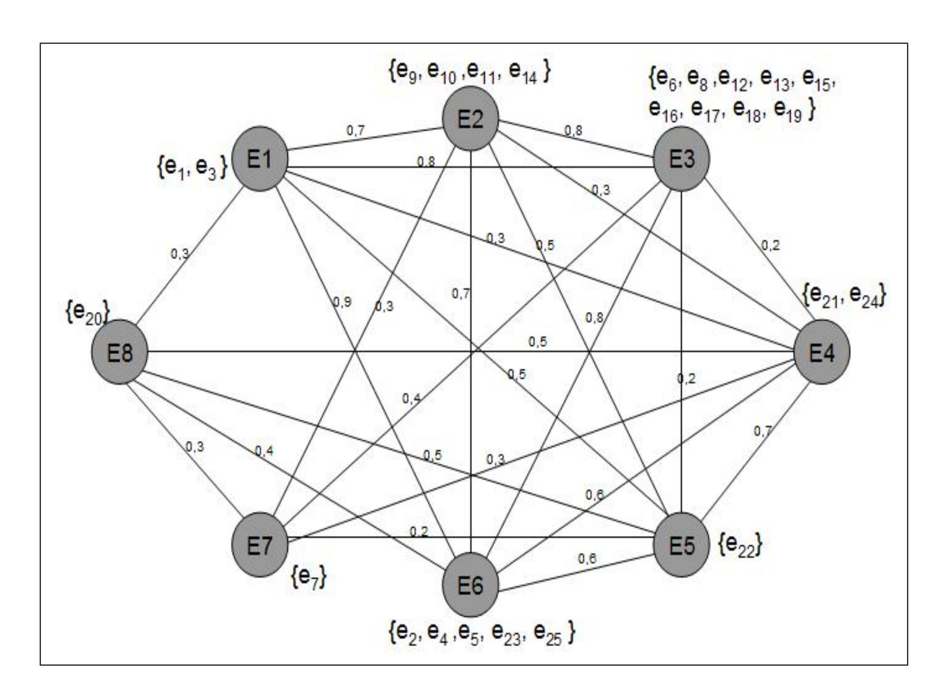


FIGURE 8.1 – Résultat du positionnement automatique des 25 entreprises sur le GCA

activités établi par l'expert. Il prend en compte non seulement la quantité d'information perdue, mais aussi la qualité d'information. La quantité d'information, c'est le nombre d'arcs éliminés. La qualité d'information est donnée par le degré de complémentarité. L'algorithme regroupe les entreprises en petits réseaux, en éliminant le moins d'arcs possibles parmi les moins significatifs (de poids faible).

Nous appliquons l'algorithme avec un pas de 0.1. Le tableau 8.1 récapitule les différentes étapes par lesquelles passe l'algorithme et les solutions données à chaque passage. I est un indicateur de qualité (noté I par la suite) de la solution de décomposition du graphe de complémentarité. Cet indicateur nous permet d'évaluer et de quantifier l'information perdue lors d'une décomposition. Si la perte d'information est trop grande, cela signifie que trop d'arcs ont été enlevés, et donc des liens de types « Réseau Proactif »ou de type « Firme »peuvent avoir été négligés. Cet indicateur se calcule d'une manière simple, par la somme des pondérations des arcs enlevés divisée par la somme totale des arcs du graphe. Plus il se rapproche de zéro, meilleure est la solution obtenue. Enfin ceci permet de savoir à quelle étape d'itération le procédé devra s'arrêter. A chaque passage (augmentation de la contrainte sur les arcs à éliminer d'un « pas »), l'algorithme donne l'ensemble des sous-groupes obtenus, la qualité de la solution (I) et l'ensemble des arcs éliminés. Un choix de trois valeurs de I correspondant à trois solutions différentes (si elles existent), permet de représenter les trois niveaux de l'intensité de la coopération de type Réseau Proactif. Les sous-groupes obtenus pour le I le plus élevé contiendront des entreprises fortement complémentaires, donc une coopération de forte intensité. Par exemple après

Etapes	Arc(k)	Arcs éliminés	I	Sous-groupe formés	Qualité
1	0.1	∅	0	∅	Faible
2	0.2	{E7, E5} {E5, E3} {E3, E4}	0.05	∅	Faible
3	0.3	{E1, E4} {E1, E8} {E2, E4} {E2, E7} {E4, E7} {E8, E7}	0.2	∅	Faible
4	0.4	{E8, E6} {E7, E3}	0.27	{E7} {E1, E2, E3, E4, E5, E6, E8}	Moyenne
5	0.5	{E1, E5} {E5, E2} {E8, E4} {E8, E5}	0.44	{E7} {E8} {E1, E2, E3, E4, E5, E6}	Bonne
6	0.6	{E4, E6} {E5, E6}	0.54	{E7} {E8} {E5, E4} {E1, E2, E3, E6}	Bonne
7	0.7	{E1, E2} {E2, E6} {E5, E4}	0.72	{E7} {E8} {E5} {E4} {E1, E2, E3, E6}	Bonne
8	0.8	{E1, E3} {E2, E3} {E3, E6}	0.92	{E7} {E8} {E5} {E4} {E1, E6} {E3} {E2}	Bonne
9	0.9	{E1, E6}	1	{E7} {E8} {E5} {E4} {E1} {E6} {E3} {E2}	Bonne

TABLE 8.1 – Construction des groupes d'entreprises

six itérations les sous-groupes suivants sont obtenus avec $I = 0.54$ de qualité :
G1= E7 ; G2= E8 ; G3= E5, E4 ; G4= E1, E2, E3, E6
L'entreprise E7 se retrouve dans les différentes étapes toute seule. c'est-à-dire avec une très faible intensité de coopération : ce qui valide les résultats trouvés par le système puisque E7 s'est retrouvée avec un score nul par toutes les fonctions de calcul de similarité.
Les entreprises E4 et E5 sont coordonnées dans une logique de coopération de type Réseau Proactif avec une intensité moyenne avec une perte d'information de $I = 0.54$. Tandis que les entreprises E1 et E6 devraient se coordonner dans une logique de Réseau Proactif avec une forte intensité de coopération, avec une perte d'information de $I = 0.92$. Selon le type de réseau que l'utilisateur veut construire. Il choisit l'étape à laquelle il s'arrête en fonction de l'indicateur de qualité I.

8.2.2 Limites de ces premiers résultats

Le test de faisabilité de l'application d'aide à la décision sur les résultats de SEI-1 a été bien validé et présente des solutions assez acceptables. L'application concrète de la méthode a supposé certaines adaptations liées à la manière dont étaient présentées les données (évaluation normalisée de degré de complémentarité, complémentarité symétrique). Les résultats positifs de ce test ouvrent la porte des réflexions pour l'amélioration de la méthode d'aide à la décision, mais cela sort du cadre de l'étude actuelle pour rester comme perspective pour des travaux futurs.

Dans la deuxième partie de cette thèse, nous avons présenté une contribution (SEI-1) qui constitue un premier système automatique d'extraction d'information sur la détection des secteurs d'activités des entreprises à partir de leurs sites web. Elle est basée sur des méthodes et des outils de recherche d'information. Les mesures de similarité utilisées s'appuient sur les indicateurs standards de la RI (Précision et Rappel) et montrent une assez bonne performance des réponses.

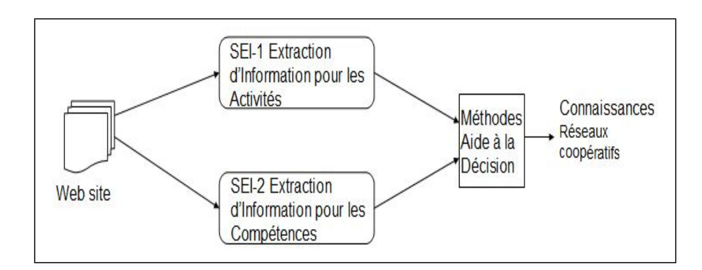


FIGURE 8.2 – Deux systèmes d'extraction d'information pour les entreprises

Cependant la complémentarité des activités des entreprises est insuffisante pour regrouper correctement les entreprises d'un même réseau de coopération. C'est pourquoi nous avons besoin du second système d'extraction concernant cette fois les compétences des entreprises (SEI-2) pour avoir une meilleure décision.
(SEI-2) est une question plus complexe à résoudre, qui nécessite le recours à des techniques d'extraction plus avancées : Analyse de texte, traitement de la langue naturelle et construction et utilisation des ontologies du domaine concernée. La partie suivante porte sur cette contribution (SEI-2).

Partie 3 : Extraction Automatique des Compétences d'Entreprises

Dans cette partie, nous présentons un deuxième enjeu de la thèse, qui consiste à étudier la capacité de construire des ressources sémantiques structurées propres au domaine métier et de les utiliser dans un processus d'extraction d'information. L'identification des compétences d'entreprises s'est avérée comme un deuxième facteur clé pour une aide à la décision en vue de construire des réseaux d'entreprises. L'approche d'extraction des compétences que nous adoptons est basée sur une chaîne de traitement des textes pour l'extraction d'information à l'aide de la construction d'une ontologie propre au domaine métier et de patrons lexico-syntaxiques. Cette contribution d'extraction des informations sur les compétences a donné naissance au système UNICOMP dont la performance est étudiée.
Cette partie est composée de quatre chapitres. Le premier chapitre fournit des éléments d'état de l'art sur la notion de compétence en génie industriel et les différents travaux sur la gestion de compétences dans les réseaux d'entreprises pour finir avec une description générale de notre approche d'extraction. Cette approche nous a permis de construire le système UNICOMP qui est décrit avec son architecture et ses différents modules. Dans un deuxième chapitre, nous mettons l'accent sur l'ingénierie d'une ontologie de domaine métier abordé, appelée « Ontologie des Traces des Compétences ». Le troisième chapitre de cette partie décrit le mécanisme d'extraction qui utilise l'ontologie du domaine et les patrons syntaxiques pour effectuer des inférences et répondre à des requêtes sur la spécification d'une trace des compétences d'une telle entreprise. Cette synthèse permet de positionner une contribution spécifique concernant l'extraction d'information sur les compétences d'entreprises qui constitue le cœur de cette partie. Enfin, le quatrième chapitre est consacré à l'étude des performances de ce système d'extraction (que nous désignerons par SEI-2).

Chapitre 9

Besoin d'extraction

9.1 Introduction

L'importance et le rôle des compétences dans l'évolution des performances industrielles [25] a fait de ce concept un axe important de recherche touchant à plusieurs disciplines (la sociologie, la science de management, l'informatique, etc). Aujourd'hui, les entreprises sont conscientes que ce sont leurs compétences qui induisent leurs performances pour survivre à une concurrence de plus en plus ardue. Dans une entreprise, les équipements matériels évoluent, les techniques et les méthodes de travail se renouvellent et les personnels se succèdent, mais les patrimoines durables reposent sur les savoirs et les savoir-faire.
Ce chapitre présente la problématique de l'extraction d'information sur les compétences d'entreprises en commençant par détailler la notion de compétence et de gestion de compétences du point de vue génie industriel (GI). Dans un deuxième temps, les méthodes et les outils d'extraction orientés compétences sont exposés.

9.2 "Compétence" en Génie Industriel

9.2.1 Définition de la compétence

Plusieurs définitions de la notion de compétence existent [22] [105] [68]. Un état de l'art exhaustif sur cette définition a été présenté dans [21]. C'est une notion pluridisciplinaire abordée selon différents points de vue. Bien que nous pouvons se référer à des travaux d'économie industrielle, de sociologie, de management des organisations etc. Nous présentons ici les définitions les plus proches de notre étude.

Le boref [22] : La compétence est un savoir-agir reconnu, un savoir-agir responsable et validé. C'est la validation qui rend compétente une façon d'agir. La compétence est une construction : c'est le résultat d'une combinaison pertinente entre plusieurs ressources (incorporées et environnementales). Ces dernières regroupent les capacités, les aptitudes, la formation et l'expérience (endogènes) ainsi que des réseaux relationnels, documentaires, d'expertise et d'outils de proximité (exogènes).

Charles-Henri Amherdt [4](compétence collective) : La compétence collective est l'ensemble des savoir-agir (*hard/soft skills and competences*) qui émergent d'une équipe de travail combinant des ressources endogènes de chacun des membres,

des ressources exogènes de chacun des membres et créant des nouvelles compétences issues de combinaisons synergiques de ressources.

Xavier Boucher [25] (macro-compétence) : Une macro-compétence est une agrégation de compétences collectives et individuelles qui permet de décrire de manière macroscopique le potentiel interne de compétence dont dispose une entreprise pour réaliser l'ensemble des activités nécessaires à sa production de biens et de services.
La notion de macro-compétence permet de décrire globalement le métier d'une entreprise, en tant que système global, et d'expliciter son positionnement stratégique basé sur la gestion de noyau de compétences "*core competence*".

Farouk Belkadi [14] : la compétence est la mobilisation d'un ensemble de savoirs hétérogènes, aboutissant à la production d'une performance reconnue, par rapport à un environnement donné et dans le cadre d'une activité finalisée.

Dans le cadre de notre étude, on utilise la notion de compétence pour décrire une vue interne de l'entreprise. La compétence nous permet de décrire les ressources et les capacités organisationnelles internes à l'entreprise, déployées pour réaliser ses activités. Pour identifier les compétences, nous nous intéressons notamment aux savoir-faire et aux expertises techniques, aux ressources techniques particulières, au savoir faire et aux expertises organisationnelles.

9.2.2 Gestion des compétences

La gestion de la compétence est un levier important pour la performance de la production et pour la coopération entre les entreprises. Avec une production croissante après les années 90, la littérature scientifique a proposé un grand nombre de contributions entendant caractériser mieux la notion de compétence.
La complexité des situations professionnelles, l'organisation du travail par réseau et le management par projet sont les facteurs qui ont poussé la réflexion sur la compétence collective. En conséquence, l'existence d'un processus de gestion des compétences dans une entreprise devient indispensable. Celui-ci a pour objectif d'améliorer les performances de l'entreprise par le déploiement efficace des compétences mobilisées dans les processus et les activités de l'entreprise.
Les travaux portant sur la gestion des compétences considèrent l'entreprise selon deux points de vue complémentaires [25] :

- L'entreprise comme un système de production de biens et de services : sa performance réside dans ce cas dans la maîtrise de ses processus de réalisation des valeurs.

- L'entreprise comme un système de production de connaissances et de compétences : sa compétitivité se fonde alors sur la maîtrise des processus de

capitalisation des connaissances et de développement des compétences.
Beaucoup d'approches d'aide à la décision appliquent la gestion de compétences dans les systèmes d'information des petites et moyennes entreprises [53]. Un état de l'art exhaustif sur l'intégration du concept de compétence dans la gestion industrielle est présenté dans [24]. Dans la gestion de la production, Franchini [59] utilise une méthode multicritère pour la gestion de compétences tout au long du processus de production. Grabot [66] inclut le paramètre de la compétence dans la planification et l'optimisation de la production. Pour sa part, Startman [152] analyse la compétence de l'entreprise pour la bonne performance de l'ERP (Entreprise Ressource Planning). Par ailleurs, différentes approches dans le cadre de la création et la gestion des Organisations Virtuelles (VO) montrent et structurent le besoin indispensable de méthodes et de systèmes d'aide à la décision basés sur la gestion de la compétence [36] [35] [54].

9.2.3 La gestion des compétences dans les réseaux d'entreprises

Dans le contexte de la coopération inter-entreprises, la gestion des compétences est prise en compte dans l'entreprise distribuée du fait que le contexte de coopération a un effet synergique sur le développement des compétences. L'amélioration de la performance globale de l'ensemble des partenaires d'un réseau d'entreprises, et notamment l'amélioration de la chaîne logistique dépend de la manière de gérer les compétences dans ce réseau.
Les travaux de [31] portent sur une modélisation des relations de coordination dans les réseaux d'entreprises. Il s'agit d'analyser les modes de coordination inter-entreprises au sein d'un réseau, à partir de la similarité des compétences et de la complémentarité des activités du réseau. Les compétences sont donc définies comme "l'aptitude à assurer la mise en œuvre coordonnée de ressources, de manière à atteindre les objectifs de l'entreprise". Des compétences seront qualifiées de similaires si elles correspondent à un même métier (mécanique, plasturgie etc). En effet, pour des raisons d'efficacité, les entreprises ont tendance à se centrer sur un noyau de compétences (core competencies). Ce qui définit leur métier de base. Les indicateurs de similarité des compétences sont calculés en utilisant la théorie des sous-ensembles flous : l'éloignement des champs des compétences de deux entreprises E1 et E2 du réseau est quantifié par le calcul de la distance de Hamming développée dans [25]. Pour l'ensemble des entreprises prises deux à deux, il est alors possible de calculer une matrice symétrique, traitée par la méthode d'analyse en composantes principales (ACP). Cette matrice permet de pouvoir visualiser un nuage de p entreprises dans un plan. Ce travail de quantification permettra de repérer les ensembles d'entreprises les plus proches sur le plan des compétences, d'aider le pilotage du réseau d'entreprises.
Pour faciliter la coopération, ces organisations ont besoin d'une infrastructure leur permettant de partager des documents, de travailler et de communiquer ensemble sans contraintes géographiques. Dans le domaine des organismes virtuels (Virtual Organizations, VO), il existe un besoin significatif des systèmes d'aide à la déci-

sion de la gestion des réseaux productifs flexibles, qui fait appel implicitement à la gestion des compétences. Des organisations (entreprises virtuelles) entendant développer leur collaboration comme un mode gestionnaire sont encore confrontées à un manque de méthodes et d'outils pour la technologie des structures coopératives agiles [24] [84] [94]. L'aide à la décision est exigée aux niveaux stratégiques et tactiques de la gestion : les systèmes de gestion de la collaboration exigent la normalisation des plateformes dédiées à cette tâche [38] et pour soutenir la conception et la création des organismes virtuels [158]. Pour aider les décideurs, plusieurs travaux de recherches ont été développés sur la formalisation des données caractéristiques des associés potentiels d'une organisation gérée en réseau et sur des mécanismes d'extraction d'information pour soutenir la prise de décision [131] [55] [37].

9.2.4 Méthodes utilisées pour l'extraction et la gestion de compétences

Les travaux de [31] sur l'entreprise virtuelle, ont proposé des méthodes et des outils d'aide à la décision pour la construction de réseaux d'entreprises basés sur la collecte et le traitement des données les concernant (les trois compétences clefs qui caractérisent le mieux l'entreprise, le niveau de maitrise de ces compétences, les personnels qui caractérisent le mieux les compétences, etc). Ces données sont collectées manuellement à partir d'un questionnaire rempli par les dirigeants d'entreprises. Il s'avère que ces derniers ne sont pas toujours collaboratifs et disposés à fournir l'information pertinente.

La plupart des travaux sur les techniques d'acquisition des compétences se concentrent sur l'analyse des textes présentant des données homogènes et structurées qui décrivent le concept de compétence dans l'entreprise ou l'organisation concernée. Ces textes sont soit des documents décrivant les compétences de l'entreprise, soit des interviews (par mail, ou oralement) faites avec les experts des entreprises pour décrire les compétences. Une fois ces données récoltées selon la structure définie, des techniques de *Text mining* sont appliquées sur le texte pour valider l'existence de telles compétences selon une description logique.

Blanchard [19] et Laukkanen [101] emploient quelques "règles expertes" basées sur les similitudes entre la définition des compétences, comme par exemple, "si les compétences Co1 et Co2 sont semblables, si un individu a acquis Co1 alors il a acquis la Co2". Pour illuster cette règle, un exemple réel peut concerner les compétences Java et C++ qui peuvent être considérées comme semblables. Si quelqu'un a la compétence Java on peut lui associer aussi la compétence C++.

[154] propose d'autres genres de "règles expertes" basées sur l'expérience professionnelle individuelle : un exemple de "règle" est ("*si un individu a participé à plusieurs projets traitant Java, alors il peut être considéré comme compétent en Java*"). Dans le dernier cas, des techniques sémantiques d'annotation sont employées pour analyser des documents traitant les activités de l'entreprise. Un autre exemple est fourni par [43] où l'annotation sémantique est également employée pour annoter les documents connexes produits par l'employé. Cette technique de gestion des compétences

est essentiellement basée sur des règles construites manuellement par l'expert du domaine. Celles-ci sont à couverture limitée. Pour identifier la compétence dans les données disponibles (documents, dossiers et données), nous devons à chaque fois reconstruire une ontologie spécifique au domaine traité.
Aucune des méthodes ne présente des résultats de performance issue d'une application réelle liée à un contexte de description des compétences. Elles se basent sur l'analyse des données homogènes collectées soit manuellement, soit automatiquement. Ces méthodes exigent des entreprises qu'elles fournissent toutes les données dans une forme structurée. Or il n'est pas évident que l'entreprise soit toujours prête à fournir cette information avec la qualité et la quantité demandée.
Il est indisponsable de disposer d'une méthode automatique de collecte et de traitement de données afin d'extraire une trace synthétique des compétences. Cette méthode doit mobiliser des techniques d'extraction puissantes (Text mining, traitement de la langue naturelle) pour produire une information dépourvue de toute ambigüité.

9.2.5 Limite des outils et des méthodes standards pour notre besoin

Notre objectif porte sur l'extraction de connaissances à partir des sites web. Ces connaissances sont des traces synthétiques des compétences que possède l'entreprise et qu'elle décrit de manière trés variable sur son site web, en décrivant ses activités, ses savoir-faire, ses produits, son équipe, ses clients, ses collaborateurs... La finalité applicative des informations extraites est de parvenir à faire émerger des propositions de collaborations inter-entreprises, à travers une similarité entre compétences. Notre corpus, qui est ainsi constitué de pages extraites de sites web des entreprises, est extrêmement hétérogène et complexe. La tâche d'extraction est rendue plus ardue par le fait qu'il n'existe pas de ressources sémantiques constituant des points de départ pour appliquer des méthodes standard d'extraction d'information (indexation sémantique).
Le site web d'une entreprise est caractérisé par une diversité des informations, à savoir des publicités, des descriptifs de produits, des informations sur les activités, l'équipe de l'entreprise... C'est un document non structuré qui comporte beaucoup de bruit vis-à-vis de notre besoin (repérer des informations sur les compétences de l'entreprise). Dans un document non structuré, par une simple recherche, nous ne pouvons pas savoir si un mot (ou une expression décrivant la compétence) est présent ou non avec une importance particulière. Alors que dans un document structuré, nous pouvons connaître avec une précision relativement fine le degré d'importance de chaque mot dans le texte. Pour extraire une compétence d'un texte, il ne suffit pas de détecter la présence ou non d'un mot. Cela nécessite une analyse contextuelle plus fine parce que la notion de compétence est liée à un réseau de concepts compliqué (par ses hyponymies, synonymies) qui dépend le plus souvent du contexte d'utilisation.
Cette difficulté de l'analyse des sites web, pour extraire les compétences d'entre-

prises, nous a conduit à nous approprier de nombreuses techniques informatiques : *Text mining*, patrons d'extraction, traitement de la langue naturelle, ontologie du domaine, pour les intégrer au sein d'une approche de traitement et d'extraction originale qui répond à notre besoin.

9.3 Notre approche d'extraction des compétences

Notre approche est basée sur l'utilisation d'une ontologie qui décrit le domaine de la compétence. Cette ontologie a été créée selon une méthodologie rigoureuse pour assurer la couverture du domaine. Ainsi représenter des relations complexes et opérationnelles à l'intérieur d'un réseau sémantique permet de rendre opérationnelle la recherche et l'extraction d'information [1]. Nous n'avons pas pu utiliser les ontologies existantes qui modélisent la notion de compétence parce qu'elles ne sont pas créées selon la description et la structuration que nous désirons. Cette nécessité pragmatique nous a poussé à concevoir et à créer notre propre ontologie appelée " Ontologie des Traces de Compétences des Entreprises ". La description de l'ontologie et sa méthode de création font l'objet du chapitre suivant.

9.3.1 Exemple de difficultés à traiter

La compétence est une notion implicite qui se manifeste dans l'activité de l'entreprise, ses produits, ses méthodes et ses outils, sa présence sur le marché et ses différentes ramifications.
"Forts de notre savoir-faire sur les machines MMAG, nous avons mis au point une gamme d'appareils haute vitesse sur paliers lisses adaptés aux exigences les plus importantes dans le domaine de la transmission de puissance mécanique" (texte extrait du site web d'une entreprise).
Forts de notre savoir faire sur les machines MMAG : comment identifier que les machines MMAG sont des produits et non pas des machines de production ?
Appareils : comment identifier que ce sont des produits et non pas des appareils de production ou de mesure ?
Exigences : cela donne des informations sur les performances, la qualité des produits. Mais comment le classifier dans cette catégorie ?
Une simple recherche des mots clés ou des concepts de l'ontologie du domaine est loin de répondre à ces questions d'ambigüité. Ce qui nécessite une analyse fine du texte, du contexte où apparait le concept de compétence, pour savoir le classifier dans la bonne classe de compétence.

1. Un site web d'une entreprise relève plus de vocabulaire publicitaire que des termes réels dont la sémantique peut identifier une compétence. C'est pourquoi on cherche à repérer des briques d'information plutôt qu'à fournir un sens complet

9.3.2 Les activités ne sont pas les compétences

"On installe, on forme, on conseille ... " extrait d'un site web d'entreprise.
Dans cet exemple d'extrait du site web, on ne voit pas ce qu'est la compétence qui est utilisée. C'est plutôt une activité et un savoir faire.
La compétence est mobilisée dans les activités. Mais la compétence ne désigne pas les activités. Au contraire, elle désigne les ressources internes à l'acteur lui permettant d'intervenir dans les activités de l'entreprise, ainsi que l'aptitude de l'acteur à mobiliser ces ressources. Pour identifier les compétences, on ne cherche pas à identifier les activités mais les ressources internes. Ces ressources internes concernent une vue structurelle de la compétence. L'aptitude de l'acteur désigne plutôt une vue fonctionnelle de la compétence. Nous nous intéressons uniquement à la vue structurelle parce qu'un site web ne donnera jamais assez d'information pour analyser les aptitudes.
Notre définition et notre modélisation de la compétence doivent rendre opérationnelle la recherche d'information. Pour extraire les informations sur les compétences, on doit faire communiquer des ontologies, (Une ontologie du domaine qui se compose d'une ontologie générique et d'une ontologie métier, voir chapitre suivant, section 3), entre elles à travers des concepts qui mobilisent la notion de la compétence d'une entreprise et à travers des instances de la compétence implicitement décrite dans le corpus. C'est cette double communication (ontologie-ontologie et ontologie-corpus) qui va créer l'intelligence dans l'extraction de l'information.

9.3.3 Le système UNICOMP

Le système que nous avons conçu, UNICOMP (UNItex COMPétence) (SEI-2) est un système dédié à l'extraction des traces de compétences des entreprises à partir de leur site web. Il prend en entrée le site web de l'entreprise et une ontologie générale décrivant toutes les compétences des entreprises (La description de l'ontologie et sa méthode de création font l'objet du chapitre suivant)(figure 9.1). Cette ontologie est structurée sous la forme de classes conceptuelles abstraites, de classes concrètes et des instances de chaque classe. En sortie UNICOMP fournit à la base des informations qui circulent sur le site web et l'ensemble des concepts du domaine de la compétence, et une liste des classes activées (sous-arbre de l'ontologie) qui valident l'existence d'un certain type de compétence.

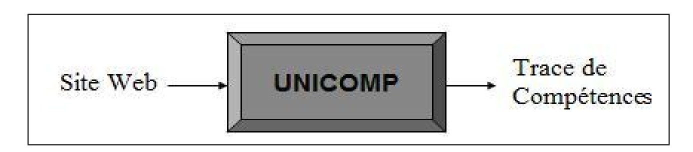


FIGURE 9.1 – Le système UNICOMP

Pour concevoir le système UNICOMP, nous avons commencé à travailler sur un protocole expérimental, qui met en place certain nombre de fragments de texte des

entreprises, choisis manuellement, comme porteurs d'une information pertinente et l'ontologie du domaine. Ces données étaient formalisées sous la forme d'une matrice qui prend en ligne le texte et en colonne les différentes classes de l'ontologie. Cette matrice était fournie aux experts, et nous leur demandons de déterminer pour chaque texte (mot, expression, phrase, paragraphe) quelles classes pouvaient, être activées selon eux. Le but de ce protocole expérimental était d'analyser le comportement de l'expert, et de comprendre sur quoi il se basait pour activer une classe de compétence. Grâce à cette expérience manuelle, nous avons pu mettre en place le système UNICOMP, basé sur une approche qui cherche à reproduire le comportement de l'expert lors de l'identification des compétences à partir d'un site web d'une entreprise :

1. Il cherche des termes de référence (*marqueurs*) décrivant la notion de compétence.
2. A partir de ces termes, il identifie des "passages délimités" qui contiennent de l'information pertinente autour des termes.
3. Ensuite, il interprète ces passages pour identifier quelles sont "les classes de compétences" effectivement trouvées (activées) dans le texte.

Cette observation du comportement humain, nous a montré que si la recherche des marqueurs est une opération facilement automatisable, par contre l'identification des passages et leur interprétation est plus délicate. Nous avons remarqué que le problème de l'extraction des traces de compétences à partir du site web de l'entreprise devient un problème d'activation des classes conceptuelles de l'ontologie décrivant le domaine des compétences. Lors de cette activation, où des classes conceptuelles sont validées par une lecture manuelle de l'expert, plusieurs phénomènes d'ambigüité ont été relevés. Cette ambiguïté est surtout liée au contexte d'utilisation du terme ou du concept. La désambiguïsation humaine faite par l'expert est reproduite dans UNICOMP par le recours à la construction des schémas structurels pour chaque marqueur de concept d'une compétence. Ce sont des schémas structurels linguistiques que peut avoir le marqueur (un mot, une expression) pour localiser implicitement ou explicitement une compétence.

9.3.4 Architecture et Modules d'UNICOMP

Le système UNICOMP est un système d'extraction de traces de compétence d'une entreprise à partir de son site web. Ce système se décompose en quatre modules (Figure 9.2) représentés ci-dessous et décrits tout au long de cette partie.

9.3.4.1 Le prétraitement

Le module de prétraitement a pour tâche principale l'extraction de texte des pages HTML constituant le site web de l'entreprise et le nettoyage du texte obtenu. Ce module est commun au premier système (SEI-1) (voir partie 2).

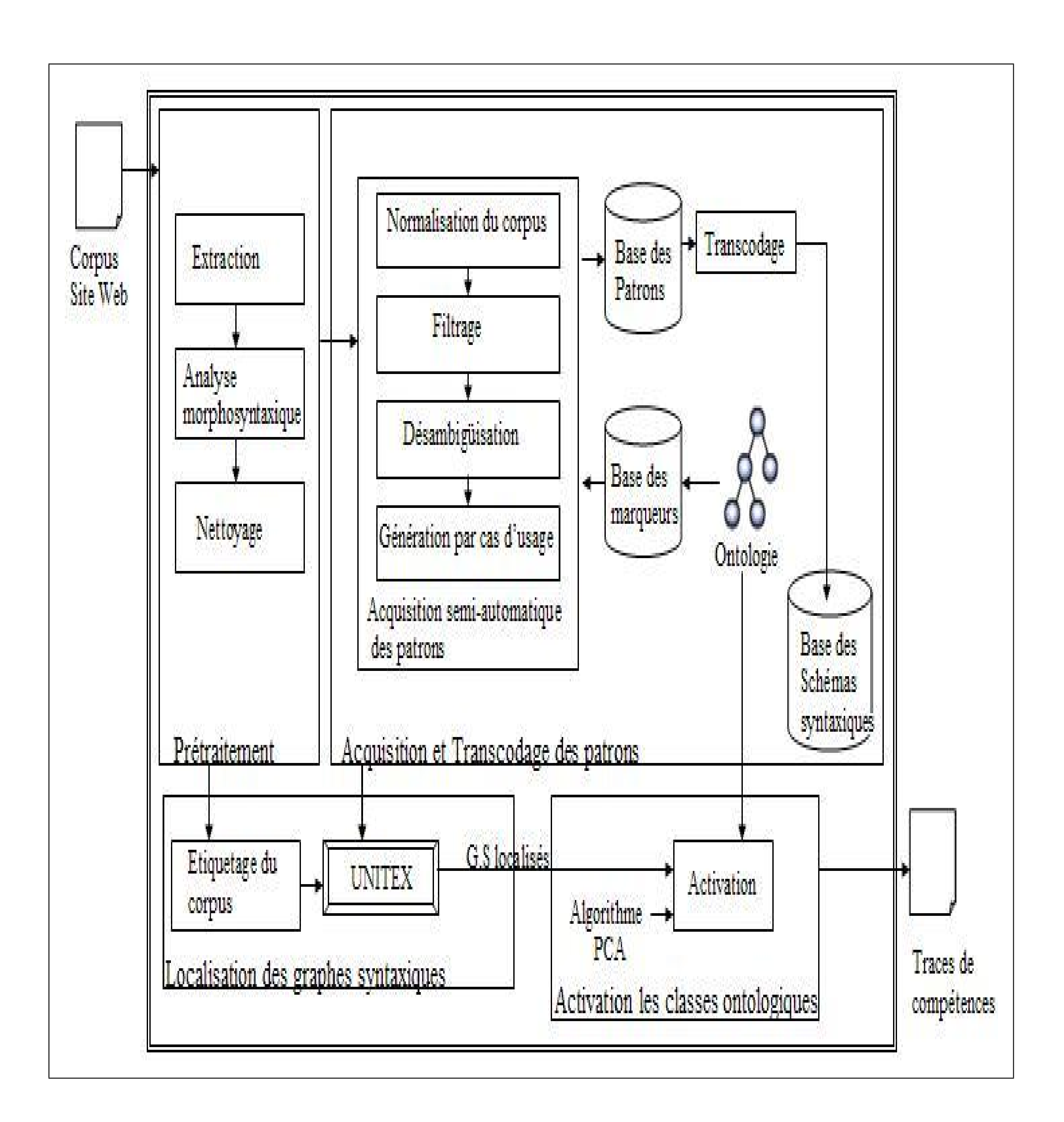


FIGURE 9.2 – Architecture du système UNICOMP

9.3.4.2 Acquisition et transcodage des patrons

L'acquisition des patrons repose sur l'observation des séquences d'informations pertinentes que véhicule le corpus. Cette observation permet de schématiser le contexte lexical et syntaxique des unités lexicales et conduit à une synthèse de ce contexte sous la forme d'un patron lexico-syntaxique. La recherche des séquences d'information pertinentes s'articule sur la recherche des marqueurs (instances de l'ontologie des traces de compétences "ontologie métier"). Une fois le marqueur repéré dans le corpus, nous pouvons extraire le plus court bloc de mots qui l'entoure et avec lequel il construit un sens non ambigu. Cette phase d'acquisition de patrons lexico-syntaxiques se compose de quatre étapes (la normalisation du corpus, le filtrage, la désambigüisation et la génération de patrons par cas d'usage) qui seront détaillées dans la section 10.4.
Le transcodage des patrons lexico-syntaxique est la transformation de ces schémas linguistiques semi-formels en schémas formels compréhensibles par le système UNITEX sur lequel nous nous basons pour faire le module de la localisation de patrons (*matching text-pattern*). Le résultat de cette transformation est un ensemble de grammaires qui représentent des phénomènes linguistiques par des réseaux de transitions récursifs (voir section 4.2).

9.3.4.3 Localisation des graphes syntaxiques

C'est le système UNITEX qui est chargé de ce module (programme *locate.exe*[2]). Celui-ci applique une grammaire à un texte et construit un fichier d'index des occurrences trouvées, leur nombre et le pourcentage d'unités reconnues dans le texte. Une faiblesse de ce module est qu'il n'effectue pas une désambigüisation syntaxique complète lors de sa recherche. Cette faiblesse est due à l'utilisation statique des dictionnaires prédéfinis. Lors de la représentation d'une phrase, tous les schémas syntaxiques sont représentés selon toutes les formes grammaticales sans tenir compte uniquement de la construction grammaticale de la phrase.
Exemple : la phrase "*Une entreprise a le produit*"[3]
Le mot produit dans cette phrase se présente comme un nom. Mais dans la représentation du dictionnaire d'UNITEX, le mot produit est représenté comme un nom (produit,.N+z1 :ms) et comme un verbe (produit,produire.V+z1 :Kms :P3s). Si nous appliquons la grammaire (figure 9.3 et 9.4) qui permet de localiser la même phrase avec le mot produit comme un verbe (ce qui n'est pas le cas) :
Unitex trouve une occurrence de cette grammaire dans le corpus. Cette ambigüité autour de l'utilisation d'Unitex a nécessité de notre part le développement d'un module qui se charge de l'étiquetage du corpus, et l'ajout d'une grammaire qui reconnait les étiquettes de TreeTagger. Ce module permet de reconstruire le corpus à

2. plusieurs paramètres peuvent être fixés lors de la recherche des occurrences : s/l/a paramètre indiquant si la recherche doit se faire en mode shortest matches (s), longest matches (l) ou all matches (a) ; i/m/r paramètres indiquant le mode d'application des transductions : mode MERGE (m) ou mode REPLACE (r), i indique que le programme ne doit pas tenir compte des transductions
3. La construction grammaticale de cette phrase est à le chapitre 5

Word Lists in C:\Users\KAFIL\Documents\Unitex\test_snt

DLF: 12 simple-word lexical entries

```
entreprise,.N+z1:fs
entreprise,entreprendre.V+z1:Kfs
le,.DET+Ddef+z1:ms
le,.PRO+PpvLE+z1:3ms
produit,.N+z1:ms
produit,produire.V+z1:Kms:P3s
une,.N+c+z1:fs
une,un.A+z2:fs
une,un.DET+Dind+z1:fs
```

FIGURE 9.3 – Dictionnaire généré par UNITEX

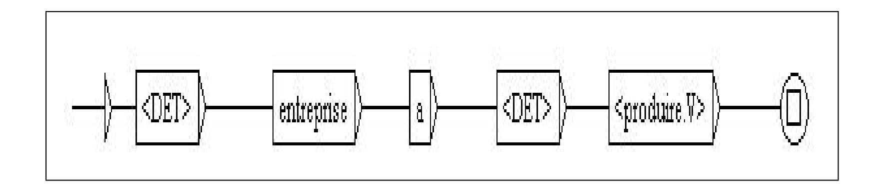


FIGURE 9.4 – Grammaire modélisant la phrase : *Une entreprise a le produit*



FIGURE 9.5 – Occurrence détectée par UNITEX

partir de l'analyse morphosyntaxique en ajoutant devant chaque mot sa catégorie grammaticale [4].
Le résultat de ce module est un ensemble validé et pertinent de patrons lexico-syntaxiques traduisant l'existence des concepts des traces de compétences dans le texte de l'entreprise. Toutefois ce résultat reste insuffisant pour déduire une trace complète et juste de compétences parce qu'une trace doit être un sous-arbre de l'ontologie de traces de compétences que l'on a construite.

9.3.4.4 Activation des classes

Le module d'activation consiste à transformer les patrons retrouvés par le module précédent en classes de concept sémantique de l'ontologie des traces de compétences. L'activation est réalisée par l'algorithme PCA (Pattern and Classes Activation) qui est basé sur un protocole bien spécifique. En effet, la seule présence d'un patron dans le texte ne suffit pas à activer une classe. Car il peut être lié à plusieurs classes ou à une ambigüité qui ne peut pas être résolue seulement par la détection du patron dans le corpus.
Comment fonctionne l'algorithme PCA ? comment est faite la désambiguïsation ? Deux questions essentielles, auxquelles nous répondons dans le chapitre 12.

9.4 Conclusion

Il est difficile d'extraire les compétences d'une entreprise à partir de son site web pour différentes raisons :
- La notion de compétence n'est pas la même pour toutes les entreprises. Elle peut avoir différentes facettes d'une entreprise à une autre.
- Le site web des entreprises n'est pas une ressource de données riche qui permet une description détaillée des compétences. C'est une source hétérogène et mal structurée qui comporte beaucoup de bruit.
- Beaucoup d'ambigüités se présentent dans la langue écrite elle-même et qui ne sont pas encore résolues informatiquement.

Le but de notre travail sur l'extraction de compétence est d'arriver à extraire non pas une carte de compétences de l'entreprise mais juste une information synthétique qui est la similarité entre deux traces de compétences.

4. c'est la bonne catégorie grammaticale qui est détectée par TreeTagger

Chapitre 10

Ontologie des traces de compétences

10.1 Introduction

La notion de compétence est une notion pluridisciplinaire abordée selon différents points de vue. Selon le but de l'étude et selon que l'analyse est issue de la sociologie, de l'économie industrielle ou du management des organisations, la définition et la caractérisation de la notion de compétence pourront être distinctes. Cette complexité de la notion de compétence rend difficile la mise au point des mécanismes d'extraction consistant à détecter une information spécifique à partir de fragments de texte (mot, expression, phrase). Dans le cadre de nos recherches, nous n'avons pas pu identifier l'existence des ressources sémantiques répondant à notre besoin de caractérisation des compétences globales et susceptibles de servir de support pour l'extraction d'information. Ce contexte, renforcé par le caractère non structuré de l'information disponible sur le web, nous a conduit à travailler sur des techniques émergentes permettant un traitement linguistique des textes : les ontologies et les patrons lexico-syntaxiques. Ce chapitre décrit la méthode suivie pour l'ingénierie et présente l'ontologie de traces de compétences des entreprises avec ses différents constituants.

10.2 Choix de méthodologie : ARCHONTE

Comme nous l'avons vu dans le chapitre 4 de l'état de l'art sur les ontologies, peu de méthodologies proposent réellement de guider l'ingénieur des connaissances pour organiser les connaissances d'un domaine et les liens entre concepts. La plupart de ces méthodes reposent sur une intuition quant à la manière de modéliser le domaine ou sur l'avis d'un expert, et excluent une possibilité de construire les concepts de l'ontologie à partir d'une réalité observée qui peut être décrite dans un langage. Le choix de la méthode d'ingénierie d'ontologie doit répondre à d'autres exigences : les textes composant le corpus ne suivent aucune structure standard ; la sémantique du vocabulaire utilisé est très liée au domaine métier (vocabulaire contextualisé) ; la structure linguistique des textes est parfois absente ; L'ensemble de ces facteurs induisent de forts risques d'ambiguïté. De plus, le choix de la méthode doit prendre en compte le fait que nous ne nous appuyons sur aucune ontologie initiale. Pour répondre à ces critères, notre choix s'est fixé sur la méthode ARCHONTE de Bachimont [9]. ARCHONTE est la méthodologie qui propose

l'approche la plus structurée et la plus complète en vue de maîtriser la spécification de la sémantique des termes, ce qui est indispensable pour traiter la problématique d'ambigüïté lors du processus ultérieur d'extraction.

Selon B. Bachimont, « *Une ontologie est une représentation linguistique et formelle des concepts d'un domaine pour un contexte applicatif. L'aspect linguistique renvoie au fait que les concepts sont tirés de la langue du domaine et doivent rester intelligibles pour les spécialistes. L'aspect formel renvoie au fait que les concepts doivent être manipulables par la machine et produire un comportement prédictible.* ». Plusieurs chercheurs [64] [72] [61] ont pu démontrer que le concept d'ontologie permet d'analyser et de traiter le savoir dans un domaine en modélisant les concepts pertinents. Les ontologies, comme ressource sémantique, sont utilisées pour aider à l'exploration de corpus. Souvent l'information pertinente se présente dans le voisinage d'un concept particulier du domaine traité, ce qui nécessite une exploration conceptuelle du texte pour la localiser. L'ontologie a notamment pour rôle de valider les entités informationnelles identifiées dans le texte. Dans notre travail, compte tenu de l'absence d'ontologie répondant réellement au besoin, il a été nécessaire d'en construire une concernant les compétences d'entreprises.

Le contenu du site web d'une entreprise est caractérisé par un vocabulaire extrêmement spécifique qui dépend directement de la réalité et du domaine de l'entreprise. Qui est-elle ? Qu'est ce qu'elle produit ? Qu'est ce qu'elle a de spécifique ? C'est un langage particulier qui n'a pas de consensus établi sur la définition des termes employés. Par exemple, sur les sites web des entreprises, le terme "haute qualité" peut se référer aux produits fabriqués par l'entreprise, comme il peut se référer aux moyens matériels ou immatériels utilisés pour réaliser un tel produit, ou à la compétence humaine qui est intervenue dans le processus de la production. Pour permettre une description efficace et dépourvue d'ambigüïté sur les compétences d'une entreprise, une modélisation qui tient compte de la réalité (ce que veut exprimer l'entreprise et la façon de le faire) et un minimum de standardisation du langage sont nécessaires.

10.2.1 Normalisation sémantique et principes différentiels

La normalisation sémantique consiste à rendre explicite le sens des expressions linguistiques du domaine. Il s'agit d'en faire des primitives du domaine. Être une primitive, c'est posséder une signification non contextuelle permettant par composition de déterminer la signification des formulations l'employant. Il faut donc identifier les notions élémentaires à partir desquelles l'ensemble des connaissances du domaine sont construites. Cette théorie attribue un sens aux termes grâce à la définition de traits sémantiques génériques et spécifiques. Ces traits permettent de fixer le cadre interprétatif, en fonction de l'objectif que s'est donné l'ingénieur des connaissances et d'obtenir une primitive exploitable. C'est une affectation des

termes aux sens qui tient compte de la variation de ces derniers dans le contexte textuel. La structuration de ces sens, en fonction des identités et des différences qu'elles partagent, permet de passer à « l'ontologie différentielle ».
Ce paradigme différentiel associe à chaque unité linguistique les unités voisines de la langue (celles qui sont utilisées en même temps qu'elle dans les contextes d'usage). Le résultat de l'application de ce paradigme différentiel est une ontologie différentielle, une structure de concepts et de relations organisée selon des principes linguistiques à partir des connaissances du domaine exprimées dans le corpus. Pour la construction de cette ontologie, B. Bachimont propose de définir quatre principes fondamentaux différentiels [9] :

- Le principe de communauté avec le père : il faut expliciter en quoi le fils est identique au père qui le subsume.
- Le principe de différence avec le père : il faut expliciter en quoi le fils est différent du père qui le subsume. Puisqu'il existe, c'est donc qu'il est distinct du père.
- Le principe de différence avec les frères : il faut expliciter la différence de la notion considérée avec chacune des notions sœurs car toute notion doit se distinguer des ses sœurs sinon il n'y aurait pas lieu de la définir.
- Le principe de communauté avec les frères : il faut expliciter la communauté entre la notion considérée et chacune des notions sœurs. Ce principe de communauté doit être différent du principe de communauté existant avec le parent.

Si nous prenons l'unité parente est *"être humain"*, les unités filles sont *homme* et *femme*. Ces unités partagent le fait d'être des humains. Mais cette propriété ne permet pas de définir en quoi sont différents les hommes et les femmes. On choisit alors comme principe de communauté la sexualité où l'on peut attribuer à *homme* le trait masculin et à *femme* le trait féminin. Ces deux traits sont mutuellement exclusifs car ce sont deux valeurs possibles d'une même propriété.
Dans les deux derniers principes, il ne faut pas seulement savoir caractériser les différences entre les notions filles mais également savoir en quoi ces notions filles sont semblables. A la fin de cette étape, on obtient une taxinomie de notions. Le processus de normalisation sémantique permet de passer d'un terme candidat à une notion dont le sens est invariable et par conséquent à une primitive représentant une connaissance du domaine à modéliser.

10.2.2 Formalisation des connaissances

La deuxième étape de la méthodologie ARCHONTE est la formalisation. C'est la définition des concepts selon une sémantique formelle et extensionnelle. C'est le passage de la dimension linguistique et interprétative de la taxinomie des termes *l'ontologie référentielle* à *l'ontologie formelle* composée de concepts dont le sens est décontextualisé. Ces concepts sont liés à un ensemble de référents dans le monde qui caractérise les connaissances du domaine. Cet ensemble est appelé l'extension du concept qui peut subir des opérations ensemblistes, telles que la réunion, l'inter-

section... qui vont permettre de composer de nouveaux sens et donc de nouveaux concepts formels. C'est l'idée derrière la notion d'engagement ontologique comme l'énonce [9] :
"Respecter le sens d'un concept, c'est s'engager à ce que lui correspond une Extension d'objets existants dans l'univers d'interprétation. Il s'agit donc bien d'un engagement ontologique, puisque c'est l'existence d'objets qui est prescrite par le sens du concept."
Cette ontologie formelle permet de définir les contraintes logiques liées à une notion, afin de les reformuler en prédicats logiques pour les intégrer de manière cohérente dans une ontologie référentielle. Cette étape permet aussi de formaliser les relations qui existent entre les concepts en définissant leur arité et les ensembles d'extensions de concepts qu'elles relient.

10.2.3 Opérationnalisation

L'opérationnalisation consiste à traduire l'ontologie référentielle dans un langage compréhensif par la machine pour manipuler les connaissances du domaine. On doit donc utiliser des mécanismes et un langage opérant sur des représentations de l'ontologie. En effet, un système informatique ne peut pas manipuler des concepts en fonction de leur interprétation sémantique. Il ne peut exploiter les concepts que sous la forme de règles formelles et d'opérations logiques (comparaison, fusion...). Ces opérations peuvent être de plusieurs sortes en fonction du formalisme de représentation choisi. C'est une définition d'une *sémantique computationnelle* pour chaque concept de l'ontologie qui sera vu comme le résultat d'un ensemble d'inférences et de calculs.
Après cette dernière étape d'opérationnalisation, l'ontologie finale peut être intégrée dans un système manipulant l'ensemble des connaissances du domaine. Elle entrera aussi dans un processus de test pour évaluer sa performance face au besoin de l'utilisateur.
Dans la section suivante, nous allons détailler la façon dont nous avons appliqué cette méthodologie pour l'ingénierie de l'ontologie de trace de compétences.

10.3 Ingénierie de notre ontologie selon la méthode ARCHONTE

Pour la construction de l'ontologie de traces des compétences des entreprises pour le domaine de la mécanique, nous avons commencé par former un corpus textuel. Celui-ci est l'ensemble des sites web des entreprises sur lesquels nous avons effectué une étape d'extraction pour générer des textes purs. Ce corpus a été soumis à une première étape d'acquisition automatique de termes. Ces termes sont destinés à une étape de normalisation proposée par ARCHONTE pour décider lesquels de ces mots ou groupes de mots sont susceptibles d'être retenus par l'expert comme des termes de l'ontologie différentielle.

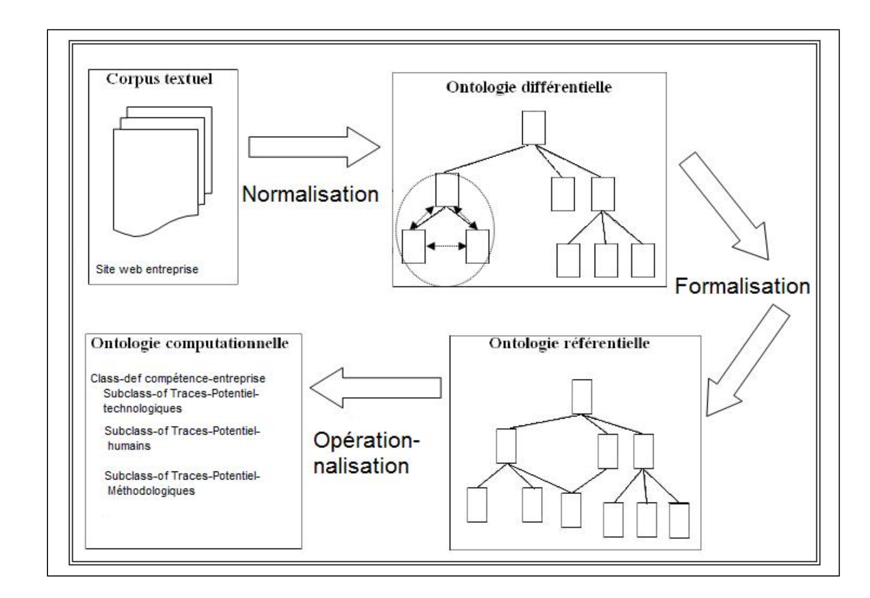


FIGURE 10.1 – La méthodologie ARCHONTE appliquée à notre corpus

En suivant les étapes d'ARCHONTE, notre objectif est de construire une ontologie concernant « les traces des compétences d'entreprises ». Une trace de compétence est une signature des ressources internes de l'entreprise, composant sa compétence. Concrètement, c'est l'ensemble des expressions types qui présentent des informations sur la compétence de l'entreprise. Il est important de souligner les ressources initiales des classes de concepts qui constitueront l'ontologie recherchée. Notre approche pragmatique consiste à se positionner dans un domaine d'activité ciblé pour les entreprises analysées (dans notre cas domaine de la mécanique). Le corpus textuel cité ci-dessus constitue une source d'analyse à partir de laquelle nous cherchons à identifier pragmatiquement des traces effectives de compétence. L'étape initiale d'acquisition automatisée a comme objectif d'aider le concepteur de l'ontologie à identifier une taxinomie initiale de termes. Ces termes ne constitueront pas directement les nœuds de l'ontologie mais sont destinés à être regroupés au sein de classes conceptuelles plus génériques. L'expert en charge de la construction de l'ontologie distingue les classes de concepts par différentiation sémantique (en référence aux paradigmes psychologique et différentiel).

Dans notre travail [78], l'ontologie de traces des compétences est composée de deux parties que nous appelons dans la suite *ontologie générique* et *ontologie métier*. L'ontologie générique permet de représenter et de modéliser le concept de traces de compétences sous forme abstraite et générique qui reste indépendante du domaine métier concerné (par exemple la mécanique dans notre cas). L'ontologie métier fournit une extension de cette ontologie, et est spécifique à un domaine métier (mécanique). Les classes de concept de l'ontologie générique sont détaillées en classes de concepts propres au métier. Ces derniers sont notamment destinés à regrouper des termes clés, susceptibles d'être identifiés dans les sites web et serviront ainsi de support à l'extraction de traces de compétences.

10.3.1 L'ontologie générique

Une ontologie ne peut être construite que dans le cadre d'un domaine précis de la connaissance, du fait que beaucoup de termes n'ont pas le même sens d'un domaine à un autre. Cette variation de sens nécessite une sémantique non ambiguë qui doit être intégrée dans l'ontologie. Délimiter rigoureusement un domaine de connaissance (dans notre cas les compétences de l'entreprise) est une tâche complexe et difficile à réaliser et nécessite une délimitation précise de l'objectif opérationnel de l'ontologie portant sur des connaissances objectives dont la sémantique puisse être exprimée rigoureusement et formellement. L'ontologie générique de haut niveau permet de spécifier les connaissances du domaine de façon indépendante du type de manipulation qui vont opérer sur celles-ci. C'est une ontologie de modélisation de la compétence qui transforme la vision d'entités de l'entreprise en une vision compétence. C'est une ontologie de domaine portant sur des concepts de haut niveau (upper-ontologies) qui offre une large possibilité de raffinement. Cette ontologie générique est construite suivant une approche descendante qui consiste à établir un modèle général pour définir la compétence d'entreprise. Ensuite, celui-ci sera raffiné en sous classes conceptuelles génériques. Pour établir ce modèle de compétences de l'entreprise, nous nous sommes appuyés sur les modèles existants en génie industriel. Nous avons cherché les principaux travaux modélisant la compétence de l'entreprise afin de proposer un modèle qui répond à notre problématique.

10.3.1.1 Modèle de Berio et Harzallah

Le modèle CRAI (Compétence-Ressources-Aspect-Individu)[17] est un modèle sémantique représentant les diverses articulations liant la compétence au contexte, aux ressources, à l'individu et à la mission. Ce modèle porte sur la modélisation de compétences, fondée sur quatre caractéristiques :

- Deux types de compétences sont distingués : les compétences acquises et celles requises.
- La compétence a des ressources structurées suivant trois catégories : savoir, savoir-faire, savoir-être.
- La compétence s'effectue dans un contexte.
- La compétence est reliée à l'accomplissement d'une ou plusieurs missions ou tâches.

CRAI modélise l'entreprise. Ce modèle sémantique représente les liens entre la compétence et toute autre entreprise modélisant des constructions : le contexte, les ressources, l'individu et les missions d'activité.

10.3.1.2 Modèle de Pépiot

Ces travaux [132] proposent de formaliser un modèle de concept de compétences destiné à être intégré dans un modèle d'organisation pour la gestion et la maintenance des compétences. Les auteurs proposent 3 types de compétences :

– Compétence unitaire : capacité à mobiliser d'une manière efficace des ressources non matérielles dans le but de répondre à une activité. Elle peut être requise par une activité ou acquise par un acteur et elle est nécessaire à l'exécution de l'activité.
– Compétence individuelle : capacité d'un acteur à combiner et à coordonner des ressources et des compétences unitaires dans le but de répondre à un objectif dans l'activité. Elle peut être requise pour le déroulement de l'activité ou acquise par l'acteur pour le déroulement de celle-ci.
– Compétence collective : capacité d'une organisation à combiner et à coordonner des ressources et des compétences unitaires.

10.3.1.3 Modèle de Hodík

Les compétences d'entreprises dans ce modèle[83] sont décrites par un ensemble de qualifications, technologies et connaissances, mais sans se rapporter à un modèle plus générique de la compétence. Basé sur l'utilisation d'un système Multi-agent, on propose un certain scénario d'utilisateur pour créer la collaboration d'affaires parmi des compagnies, avec une vue explicite sur leurs compétences internes.
Le concept de compétence d'entreprises est défini comme un ensemble de qualifications, de technologies et de savoir-faire, sans se référer à un modèle général de compétence. Les auteurs définissent un scénario d'utilisation des compétences qui est guidé par un expert (ajouter compétence, éditer compétence, supprimer compétence...) dans le cadre de la gestion de profils et de compétences d'entreprise pour la création des organisations virtuelles.

10.3.1.4 Modèle d'Yussopova

Une ontologie formalise les concepts nécessaires pour représenter et contrôler la mémoire de corporation d'une entreprise [171]. Cette ontologie est suggérée comme soutien de gestion des compétences.

10.3.1.5 Modèle de Boucher et Burlat

Le modèle qualitatif s-a-r-C proposé dans [25] spécifie le concept de compétence en tant qu'émergence de l'interaction entre trois composantes essentielles :

– *Situation* : Dans ce modèle, une « situation professionnelle »sera modélisée par les attributs : un ensemble de « problèmes caractéristiques »auxquels l'acteur est confronté, un « objectif »qui spécifie l'enjeu de la situation (ce qui lui donne un sens pour l'acteur) et le résultat à atteindre (variable observable qui permet de contrôler l'atteinte des objectifs) et un « contexte »c'est à dire un ensemble de facteurs (contrôlables) qui ont un impact sur la compétence.
– *Acteur* : Les acteurs désignent les ressources humaines de l'entreprise, qu'elles soient individuelles ou collectives (l'acteur intègre la notion de ressource immatérielle).

- *Ressource* : le concept de ressource est utilisé ici pour décrire de manière exclusive les ressources de type matériel.

10.3.1.6 Modèle d'Ermilova et Afsarmanesh

Le concept de compétence fait partie du concept du profil des VBE [54], où un ensemble de membres des organisations se mettent d'accord pour travailler et collaborer ensemble en fournissant et en partageant certaines ressources et des informations à court ou long terme. Dans cette modélisation, le concept de compétence joue un rôle important pour faire évoluer la collaboration entre ces différents membres, et faire émerger des organisations virtuelles VO. Il est modélisé pour fournir une description structurée des profils des entités de VBE qui va être utilisé pour la création des VOs. Ces compétences englobent principalement les possibilités et les capacités des entités de la VBE. La compétence est identifiée dans le cadre de la collaboration des réseaux pour former des organisations virtuelles.
La modélisation des compétences des VBEs est composée principalement des trois éléments :

- *Capability* : Cet attribut représente la liste de toutes les capacités des membres de l'organisation de VBE qui participent à l'émergence des nouvelles organisations virtuelles. Cette notion de capabilité représente un aspect important dans la constitution de la capacité du VBE. Il représente l'ensemble des processus d'activités qui peuvent être exercés et qui peuvent contribuer au développement des VOs. Les attributs principaux de cette classe sont le nom, la description, le temps d'exécution et le rendement.
- *Capacité* : Cet attribut représente les disponibilités des ressources dans les membres de l'organisation de VBE qui participent à l'émergence des nouvelles organisations virtuelles. Cette notion de capacité se réfère à la disponibilité en termes de temps et de pourcentage des ressources et des partenaires associés à cette VBE.
- *Conspicuity* : Cet attribut représente l'ensemble de documents qui peuvent indiquer la validité d'autres données de compétence fournies par les organismes. Les données de compétence qui ont été fournies par le membre de VBE peuvent être représentées par un certain nombre de documents qui peuvent ajouter différents niveaux différent de validité à leurs affirmations. Les deux sous-classes principales de l'évidence sont identifiées dans ces documents en tant que *fact-based* (par exemple certificats, récompenses, brevets) et *opinion-based* (par exemple lettres de recommandation).

10.3.1.7 Notre Modèle de Compétence d'Entreprise

Les modèles de compétence présentés ne répondent pas à notre besoin, ils ne modélisent pas tous l'entreprise selon une vue interne et externe en tenant compte des compétences individuelles et collectives.
Notre modèle de compétences des entreprises (figure 10.2) se base sur deux notions

principales : une compétence émerge comme une combinaison de capacités internes. Ces capacités elles-mêmes sont le résultat de la mobilisation de différentes ressources que possède l'entreprise. Pour raffiner ces deux notions, nous partitionnons les ressources en quatre types élémentaires : ressources humaines, technologiques, informationnelles et organisationnelles. En outre, nous distinguons deux types de capacités : capacités technologiques, se rapportant à la création de valeur ajoutée basée sur l'utilisation de ressources et de processus techniques, et capacités méthodologiques lié à la valeur ajoutée fournie par les méthodes de travail employées par une entreprise pour fournir ses produits ou ses services.

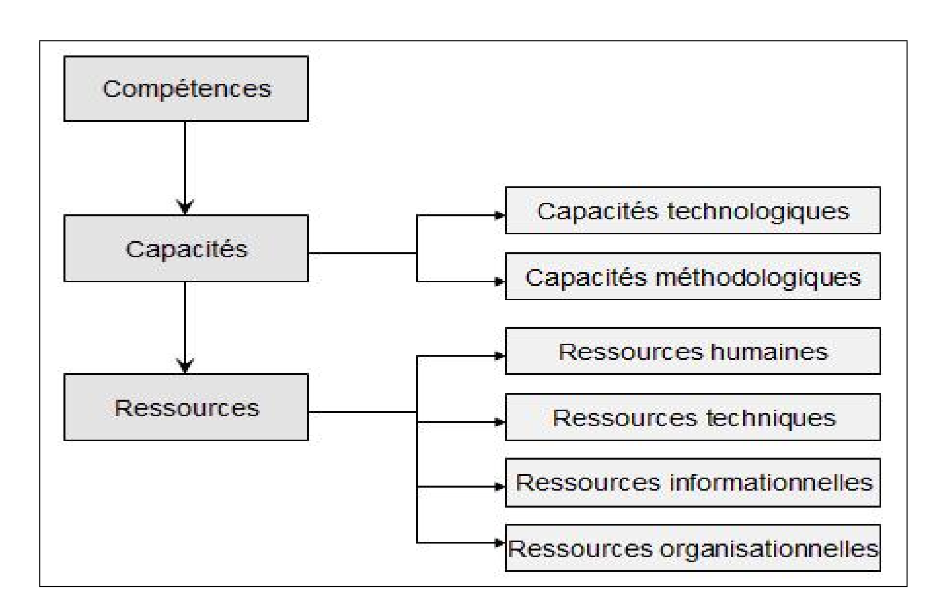


Figure 10.2 – Notre Modèle des compétences d'Entreprise

Notre modèle [78] fait référence à toutes les compétences reliées au savoir, et au savoir-faire de la technologie : équipements, procédés de production, ressources techniques... Les capacités méthodologiques regroupent les compétences reliées à l'acteur (individu ou groupe d'individus), qui reflètent leurs connaissances, leur savoir et leur savoir-faire, leur expertise et leur qualification. Les capacités méthodologiques recouvrent une double vision : en interne de l'entreprise pour modéliser son savoir organisationnel du travail, et en externe pour modéliser sa capacité de réactivité, d'écoute du client et d'adaptation à ses besoins [17] [132][83] [171] [54].
L'ontologie générique est une ontologie de modélisation de la compétence pour la modélisation de l'entreprise. Cette ontologie générique est construite suivant une approche de construction descendante qui consiste à établir un modèle général pour définir la compétence d'entreprise, le modèle qui est ensuite raffiné en sous classes conceptuelles génériques.

Dans notre ontologie générique sur les capacités, on génère des classes du concept de "trace des compétences". Ces classes sont construites à partir d'une analyse du corpus et d'une confrontation avec l'expert du domaine pour normaliser leurs

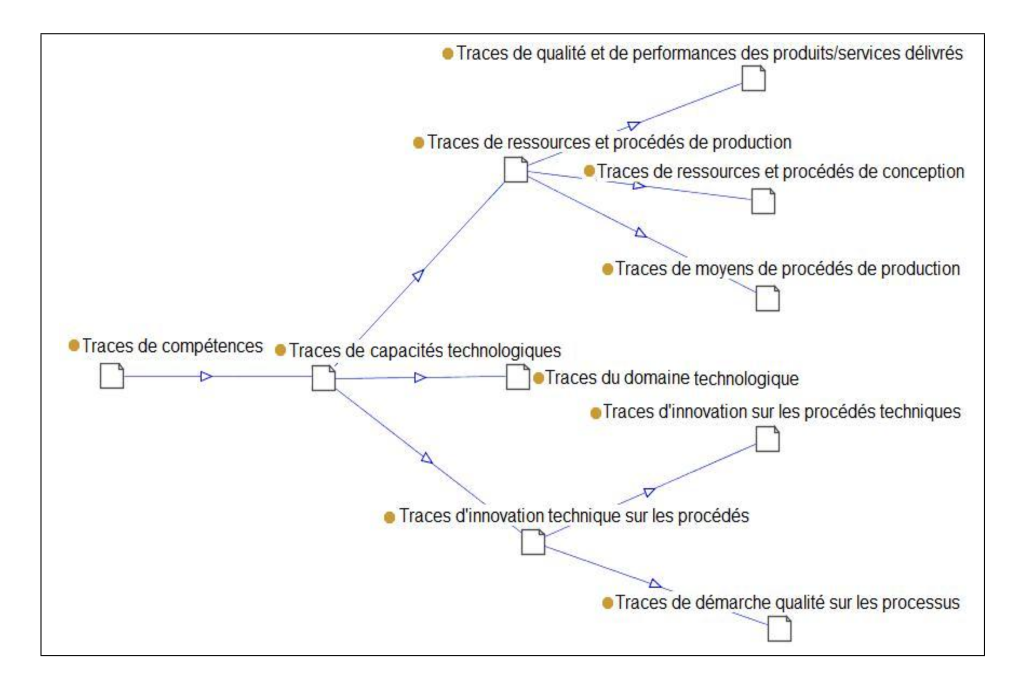


FIGURE 10.3 – Extrait d'ontologie générique

significations et garder un sens complet dépourvu de toute ambigüité. L'ontologie générique est composée de deux niveaux : un premier niveau qui manipule les concepts abstraits (traces des capacités techniques) et un deuxième niveau composée des concepts structurants (traces des ressources et processus techniques, traces du domaine technologique, etc) décrivant plus en détail les potentiels du modèle de compétence (figure 10.2) sous forme de classes conceptuelles génériques. Une classe conceptuelle peut être définie dans notre cas comme une entité qui regroupe toutes les caractéristiques sémantiques liées à une idée d'un domaine des compétence des entreprises. Cette idée est exprimée en fonction d'un terme ou d'une expression.

10.3.2 Ontologie Métier

L'ontologie métier permet la description et la classification des connaissances des domaines moins abstraits. Ces connaissances sont moins générales, et leur utilisation est beaucoup plus dépendante du domaine métier de l'entreprise. Durant la construction de cette ontologie, nous nous sommes intéressés à plusieurs classifications possibles, qui peuvent donner les mêmes instances mais selon plusieurs points de vue. Cette problématique de diversification et de choix de classification est liée à chaque fois à une nécessité pragmatique qui se résume dans la nature, la qualité et la quantité de l'information manipulée et présentée dans le corpus (site web de l'entreprise). En effet, une vision externe d'une telle classification des classes de concepts métiers de l'entreprise peut parfois contredire la classification interne et réelle liée au contenu du site web de l'entreprise. Ce choix pragmatique de la création et la classification des concepts tiennent compte aussi des nouveaux termes et concepts qui arrivent et qui devraient être inclus sans réviser toute la structure

et les définitions existantes de l'ontologie.
L'ontologie métier est l'ensemble des types de concepts qui incluent l'ensemble des marqueurs et déclencheurs candidats pour exprimer une compétence (technique, individuelle...) de l'entreprise. Un marqueur est un terme ou une expression qui permet d'introduire (déclencher) une idée liée au domaine de connaissances (compétences d'entreprise) et signaler la présence d'une compétence dans le corpus étudié. Comme il est presque impossible de dénombrer toutes les compétences des entreprises dans tous les domaines, on utilise ce type d'ontologie (ontologie métier) pour la détection de la présence (ou la forte possibilité de présence) d'une compétence spécifique dans le corpus. Par exemple, les termes marqueurs *outils, outillage* déclarent la présence d'une compétence technique. Cette compétence technique peut être aussi associée à d'autres macro-compétences. Cette ontologie est changeable et doit être construite pour partir d'un domaine à un autre. Ce qui nous assure sur la qualité aux traces de compétence extraites, vu que la décomposition de la compétence n'est pas la même d'un métier à un autre. Les classes de marqueurs (types de concepts) du domaine mécanique (outils, outillage) ne sont pas les mêmes que le domaine de l'informatique (système d'exploitation, programmation, base de données...).
Dans la suite, nous présentons un extrait de notre ontologie métier construite pour le domaine de la mécanique. Les classes de l'ontologie métier dérivent toutes d'un

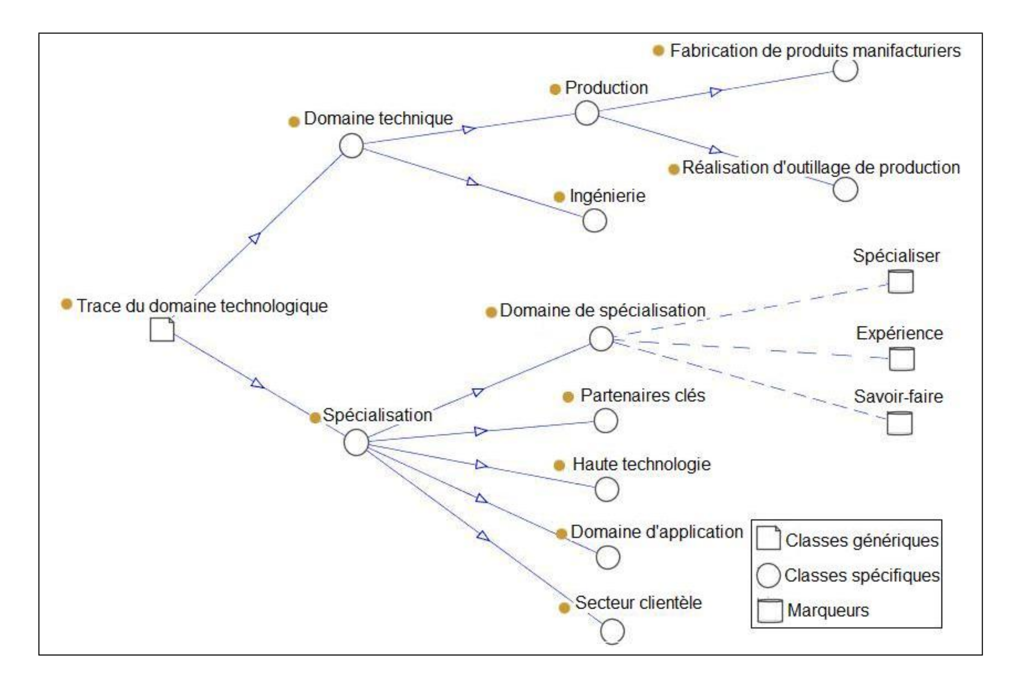


FIGURE 10.4 – Extrait d'ontologie métier

haut niveau abstrait, qui est le potentiel technologique de l'entreprise et qui appartient à l'ontologie générique. On trouve par exemple la classe technologie qui dérive trois autres classes (usinage, traitement de surface, assemblage). Nous remarquons qu'il ne s'agit pas d'une classification des compétences d'une entreprise dans le domaine de la mécanique mais plutôt d'une classification des termes et des marqueurs

qui impliquent un concept de compétence. Sur les sites web des entreprises, nous pouvons croiser par exemple les phrases ou les expressions suivantes :
Nous sommes spécialisés dans l'usinage haute vitesse.
Nous utilisons la technologie laser
Dans la première phrase le marqueur *usinage* introduit une compétence dans le domaine technologique qui est la *haute vitesse*. Dans la deuxième phrase c'est le marqueur *technologie* qui déclenche la compétence *laser*. Ces deux marqueurs sont les briques d'information que l'on cherche à retrouver dans le texte qui est diffusé sur le site web de l'entreprise. Rappelons que notre but final n'est pas l'identification détaillée de la compétence de l'entreprise (on ne cherche pas à extraire la carte de compétence de l'entreprise) mais plutôt l'identification d'une trace de compétence pour extraire une information synthétique qui indique la similarité entre deux entreprises en terme de compétence. Cette trace de compétence est identifiée à partir de la double communication Ontologie-Corpus et Ontologie-Ontologie.

10.3.3 Normalisation de l'ontologie

Une ontologie est une représentation formelle des éléments conceptuels et de leurs relations constitutifs d'un domaine de connaissances. Il ne s'agit pas de représenter et de modéliser une expertise ou des processus cognitifs des personnes. Il s'agit d'une modélisation d'un domaine qui correspond à un champ de pratique. C'est pourquoi l'étape de normalisation est primordiale dans le processus de construction de l'ontologie. C'est une normalisation linguistique qui permet un choix des termes dans un contexte de référence. Il ne suffit pas de détecter qu'un terme dans le corpus exprime une connaissance. Il faut établir laquelle et contraindre l'utilisateur à un engagement sémantique en introduisant une normalisation sémantique des termes manipulés dans l'ontologie. Pour passer des unités linguistiques extraites à des concepts ontologiques primitifs, il faut d'une part dégager la signification des unités extraites et d'autre part la déterminer suffisamment et précisément, pour définir un concept primitif possédant une signification non contextuelle. La normalisation sémantique est basée sur une sémantique différentielle qui détermine le signifié des unités linguistiques en termes de traits différentiels.
Dans notre ontologie, le processus de la normalisation est effectué en deux étapes : la première consiste à identifier automatiquement une série des termes candidats de l'ontologie avec l'outil d'indexation SMART. Ce dernier permet de proposer une liste de termes ordonnés selon leur fréquence d'apparition dans le corpus. La deuxième étape consiste à valider et à ressortir une autre série de termes en présentant des exemples de notre corpus à des experts du domaine de la mécanique pour donner une signification précise aux termes de l'ontologie. Nous avons travaillé avec des experts du domaine sur un protocol expérimental. Nous avons cherché à partir de leurs réponses quel concept générique pouvait être inclus suite à cette confrontation. Nous avons rencontré des problèmes d'ambiguité des termes, de synonymie et d'opposition, ou des difficultés de termes intra-linguistiques. Ainsi avec nos experts, les traits sémantiques qui déterminent le sens des termes (sème) ne sont pas forcé-

ment les mêmes. Cette problématique a été abordée par le recours à la méthode des juges [1]. C'est une mesure pour évaluer la cohérence des réponses des juges (experts du domaine). La fiabilité est fondée sur la corrélation ou l'analyse de la variance. Ce sont des indices qui permettent d'évaluer dans quelle mesure les avis des différents juges sont les mêmes, exprimés en écart par rapport à leur connaissance et à leur représentation mentale de l'objet qui fait référence au concept.
Pour justifier la construction de l'étape de normalisation et afin d'avoir une ontologie composée d'une structure de concepts et de relations organisée selon des principes linguistiques, nous avons travaillé sur le choix des termes comme c'est indiqué ci-dessus, pour éviter toute ambigüité de sens des termes. La structure du réseau des concepts est un arbre, nous avons travaillé sur la signification que doit posséder chaque nœud en fonction de sa position dans l'arbre (en appliquant le paradigme différentiel proposé par [9]). Cette analyse a été faite avec nos experts pour expliquer en fonction des voisins, les identités et les différences qui définissent chaque nœud, comme décrit dans la section 9.2.1 :

- Le principe de communauté avec le père
- Le principe de différence avec le père
- Le principe de différence avec les frères
- Le principe de communauté avec les frères

Pour chaque concept, ces questions ont fait l'objet de discussions entre les experts du domaine et l'analyste, pour valider ou non le choix de ce concept. Le résultat final est une ontologie différentielle basée sur la sémantique de la signification (associer à chaque concept une signification linguistique).

10.3.4 Formalisation de l'ontologie

C'est le passage de la sémantique de la signification à la sémantique de la désignation. Cette étape est cruciale pour rapprocher l'effectivité calculatoire de l'intelligibilité conceptuelle. L'objectif est de doter chaque concept de l'ontologie interprétative d'une référence. Pour justifier l'étape de la formalisation on a ajouté des propriétés à chaque concept (métaphysiques, structurants, parataxiques) qui caractérise les différentes manières de penser. Au premier niveau, on trouve des concepts très abstraits, introduits pour structurer le reste de l'ontologie. Ce niveau est inspiré du modèle conceptuel des compétences des entreprises. Ces concepts reposent sur un premier niveau de modélisation de la compétence. Le deuxième niveau contient les concepts de base du domaine que l'on utilise pour structurer les connaissances. Au troisième niveau on trouve des concepts qui servent à désigner des objets du domaine dans un monde énumératif. La formalisation de l'ontologie différentielle correspond à une instanciation des concepts précédents et le choix d'une référence respectant les contraintes fixées par le concept ainsi instancié. Par exemple dans l'ontologie métier, le concept *usinage* instancie le concept *technologie*. Ces deux derniers concepts parataxiques instancient eux même un

1. http ://www.temple.edu/sct/mmc/reliability/

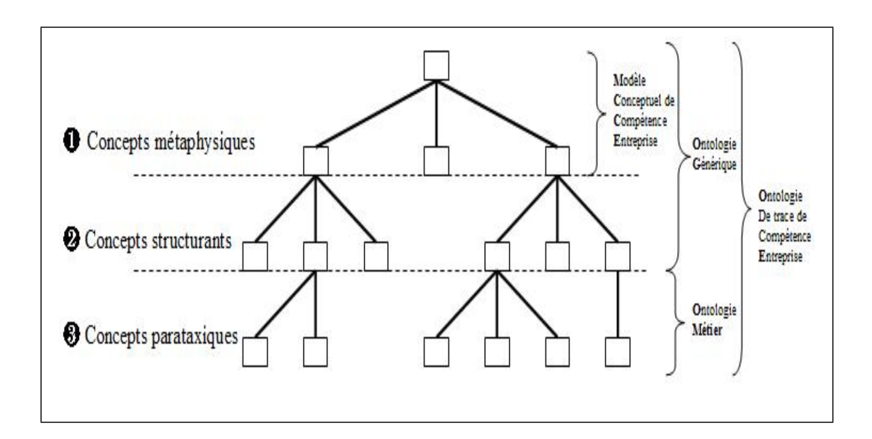


FIGURE 10.5 – Formalisation de l'ontologie différentielle

concept structurant qui est *trace du domaine technologique*.

On a ajouté dans cette phase de normalisation une relation d'association entre les concepts qui permet de lier deux concepts. Par exemple des concepts de qualités et de performance des produits peuvent décrire des traces de ressources techniques. C'est l'avantage de cette relation d'association qui permet de renvoyer à d'autres concepts, ce qui est important par la suite dans la phase de recherche et d'extraction. L'analyse de la même phrase peut conduire à détecter la présence de deux (ou plusieurs) concepts.

10.3.5 Opérationnalisation de l'ontologie

C'est l'élaboration d'une version de l'ontologie exploitable informatiquement où la signification des concepts se traduit par des calculs ou des inférences. L'ontologie computationnelle est construite dans le langage OWL (*Ontology Web Language*) avec l'outil Protégé [2]. Ce langage fournit des primitives de modélisation permettant de déclarer les ontologies et d'exprimer précisément leur sémantique. C'est l'outil formel pour contraindre la syntaxe ontologique (figure 10.6).

10.4 Conclusion

Nous avons présenté les principaux outils et méthodes de construction des ontologies parmi lesquels nous avons effectué notre choix. Nous avons construit une ontologie des traces de compétences des entreprises dans le domaine de la mécanique suivant la méthode ARCHONTE tout en respectant les différentes primitives cognitives, à partir de notre corpus (collection de site web des entreprises). Une question importante reste à résoudre : comment exploiter cette ontologie pour mener des inférences et répondre à des requêtes sur la spécification d'une trace des compétences d'une entreprise donnée ?

2. http ://protege.stanford.edu/overview/protege-owl.html

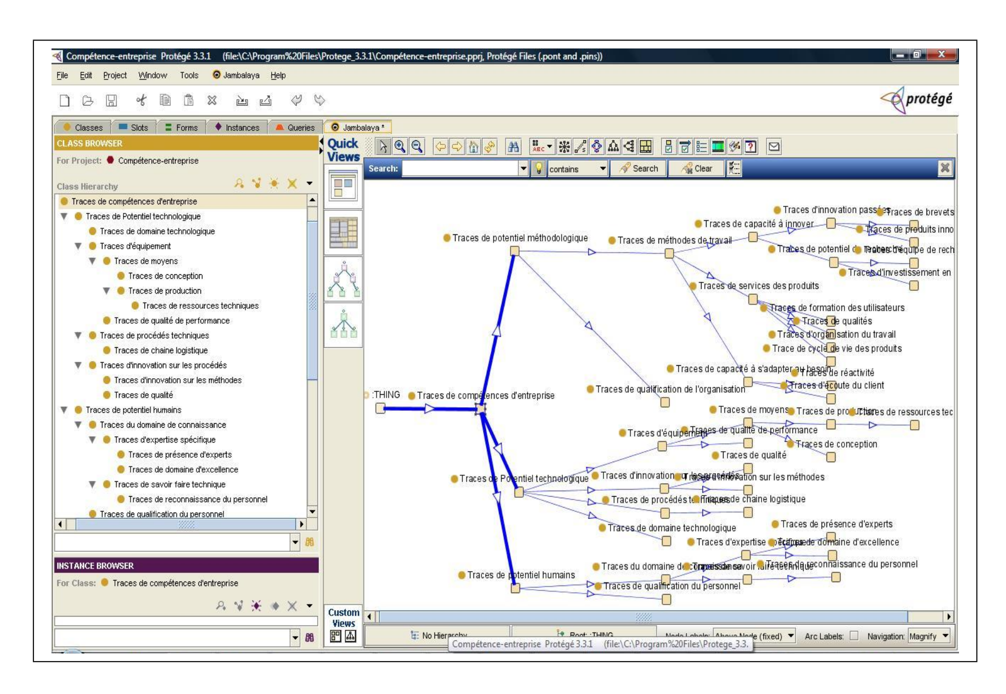


Figure 10.6 – L'ontologie computationnelle

Chapitre 11

Extraction de compétences

11.1 Présentation de l'application

Comme nous l'avons mentionné dans le chapitre 5 de l'état de l'art, le système d'extraction des traces des compétences des entreprises à partir de leur site (UNICOMP) est basé sur le système de traitement linguistique Unitex. Unitex permet de traiter un corpus textuel pour l'indexation de motifs morphosyntaxiques, la recherche d'expressions figées, la production de concordances et l'étude statistique des résultats. Un aperçu des ressources développées lors du traitement d'un texte est donné en figure 11.1 :

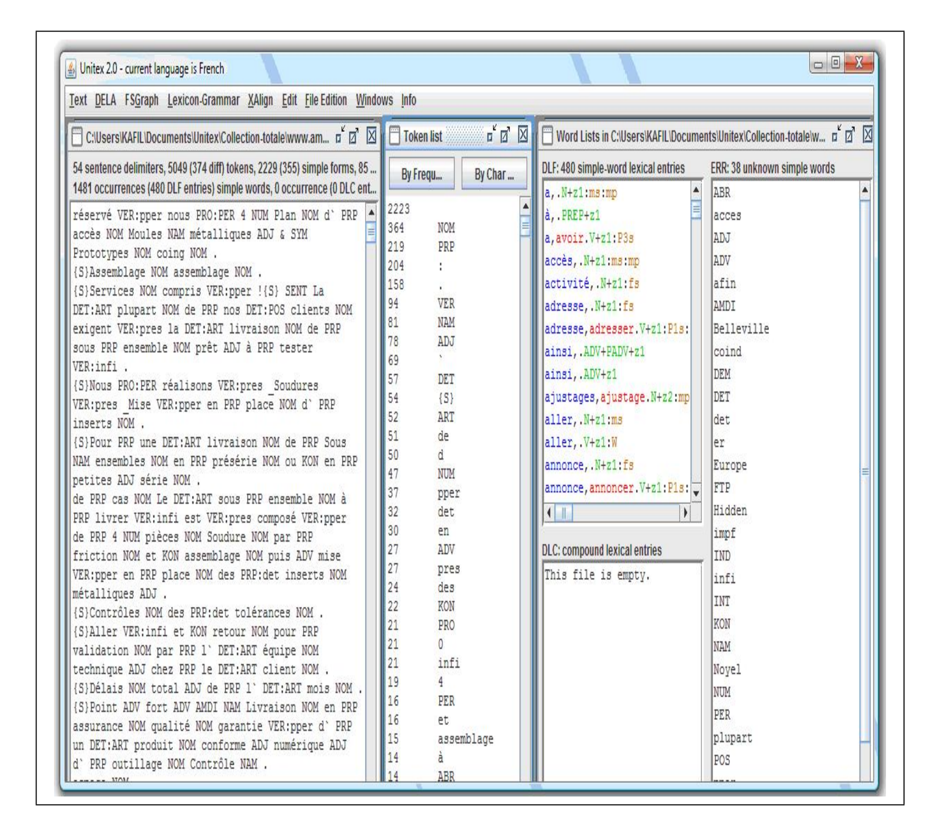


Figure 11.1 – L'application UNITEX

Le panneau à gauche présente le corpus prétraité après avoir effectué le découpage en phrases. On voit la liste de tous les *tokens* (au milieu) avec les fréquences d'apparition, ainsi que les unités linguistiques (à droite) traitées par les

dictionnaires de mots simples et de mots composés. La dernière colonne représente les unités linguistiques qui n'ont pas été retrouvées dans les dictionnaires. Unitex est utilisé comme analyseur pour effectuer un prétraitement et une lemmatisation des mots, pour ajouter des synonymes, pour détecter la négation, pour ajouter des classes sémantiques aux mots, et enfin et surtout pour l'extraction, la construction et la recherche des grammaires locales complexes.

En extraction d'information, la détection au sein d'un texte de la présence d'un concept issu d'une ontologie n'est pas une condition suffisante pour délimiter et confirmer l'information pertinente. Des phénomènes linguistiques peuvent biaiser le sens des mots et un même mot peut prendre deux sens différents selon son contexte d'utilisation. Pour lever cette ambigüité contextuelle, en complément à l'ontologie, nous aurons recours à l'utilisation de patrons linguistiques implémentés par le système UNITEX. La section suivante présente l'approche adoptée pour l'acquisition des patrons d'extraction.

11.2 Acquisition semi-automatique de patrons d'extraction

La stratégie mise en œuvre cherche d'abord à filtrer les séquences pertinentes du corpus autour de l'ensemble des marqueurs de l'ontologie, trouver des mots sémantiquement proches puis s'assurer que ceux-ci se trouvent au sein d'une structure syntaxique spécifique. Cette phase d'acquisition se compose de quatre étapes.

11.2.1 Normalisation du corpus

Cette étape consiste à remplacer les mots qui ont le même sens et qui sont pertinents pour le domaine par un seul terme ou expression indiquant le nom de la classe sémantique générale. Ainsi le terme *entreprise* peut être exprimé par différentes expressions : notre entreprise, notre société, nous, le nom de l'entreprise... Ces expressions sont remplacées par le nom de la classe sémantique qui est dans ce cas *Représentant de l'entreprise*
Des exemples issus du corpus :

- ATTAX conçoit, industrialise, et commercialise des dispositifs de fixations destinés à toutes les industries.
  Représentant entreprise conçoit, industrialise, et commercialise des dispositifs de fixations destinés à toutes les industries.
- La société MECADEX est spécialisée dans le décolletage de précision.
  Représentant entreprise est spécialisée dans le décolletage de précision.
- Fabricant des appareils pour bancs d'essais, nous pouvons prendre en charge leur conception intégralement pour toutes applications tournantes.
  Fabricant des appareils pour bancs d'essais, **Représentant entreprise** pouvons prendre en charge leur conception intégralement pour toutes applications

tournantes.
Les marqueurs sont aussi identifiés dans le corpus par une recherche automatique de leur lemme grâce à la fonction *Locate* avec une expression régulière ;
Exemple : <spécialiser> : reconnaît toutes les entrées dont la forme canonique est le mot "spécialiser".

11.2.2 Filtrage des phrases pertinentes

Le filtrage des phrases pertinentes est effectué grâce à l'ensemble des marqueurs (instances de l'ontologie des traces de compétences). Il s'agit de ne retenir que les phrases et les paragraphes qui sont potentiellement pertinentes (phrases où apparaissent les marqueurs) pour éviter à l'expert de lire tout le corpus. Ainsi les phrases présentées dans l'étape précédente sont des exemples qui ont été filtrés du corpus.

11.2.3 Identification d'exemples représentatifs

C'est la détermination, parmi les phrases filtrées, des syntagmes représentatifs. C'est l'ensemble des termes qui peuvent et/ou doivent être corrélés au marqueur pour définir un sens pertinent pour notre recherche. A cette étape, seule l'expertise humaine est capable de déterminer et d'évaluer la pertinence d'un syntagme.
L'identification des syntagmes est basée sur une analyse par ambigüité. Cette analyse consiste à chercher toutes les ambiguïtés que peuvent avoir le marqueur dans le contexte du corpus. Quelques ambiguïtés qui ont été levées du corpus :
Représentant entreprise La société Technax industrie, basée à Genas (Lyon, France), est spécialisée dans la conception et la réalisation de machines d'assemblage...
Ambiguïté : Il faut faire la différence entre conception des produits et conception des outils de production. Dans le cas de cette phrase, il s'agit de conception des outils de production.
Représentant entreprise s'est dotée de tous les moyens, techniques et humains, pour atteindre la haute performance dans l'infiniment précis.
Ambiguïté : Comment identifier que les notions de qualité/performance concernent les produits et les services ou les outils et les méthodes utilisées pour la production.
Le tableau 11.1 résume des types d'ambigüités que l'on retrouve dans des classes conceptuelles de l'ontologie.

11.2.4 Génération des variantes de patrons

Le module de génération des variantes de patrons a pour rôle d'étendre la couverture du système en proposant des structures sémantiquement équivalentes. Cette étape se base sur l'expertise humaine et sur la recherche et l'analyse d'autres exemples d'entreprises sur le web. L'ensemble des patrons constitue une bibliothèque de patrons dont nous montrons ci-dessous un extrait.
Les patrons peuvent être utilisés pour trois cas d'usage :

Classe conceptuelle	**Ambigüité**
Traces des ressources et procédés de conception	Comment identifier que ce sont des moyens et non des produits ?
Traces des ressources et procédés de production	Comment identifier que ce sont des moyens de production utilisés et non des équipements vendus à d'autres entreprises ?
Traces de qualité et de performance des produits/services	Comment identifier que ces notions de qualité/performance concernent bien les produits et les services ?
Traces d'innovation sur les procédés techniques	Comment identifier que cela concerne les procédés et non l'innovation produit ?
Traces de démarche qualité sur les processus	Certaines sous-classes requièrent une analyse linguistique pour l'extraction d'autres informations que le marqueur

TABLE 11.1 – Type d'ambigüités par classe conceptuelle

PATRONS	**USAGE**	**CLASSE A ACTIVER**
Représentant Entreprise - produire	Détection de la présence d'un Concept (DPC)	PRODUCTION - PROCEDES DE FABRICATION
Représentant Entreprise - fabrique	DPC	PRODUCTION - PROCEDES DE FABRICATION
Représentant Entreprise - verbe d'action - COD	Désambigüisation Entre Concepts (DEC)	REALISATION D'OUTILLAGE DE PRODUCTION
Représentant Entreprise - verbe d'action - COD	DEC	FABRICATION DE PRODUITS MANUFACTURIERS
Ingénierie	DPC	INGENIERIE - PROCEDES D'INGENIERIE
Haute technologie	DPC	HAUTE TECHNOLOGIE
Technologie de pointe	DPC	HAUTE TECHNOLOGIE
Technologie innovante	DPC	HAUTE TECHNOLOGIE
Représentant Entreprise - forme verbale passive incluant spécialisé - PREP-GN	Extraction d'Information Complémentaire Rattachée au Concept (EICRC)	SPECIALISATION - DOMAINE D APPLICATION
Expérience - PREP - GN	EICRC	SPECIALISATION - DOMAINE D APPLICATION

TABLE 11.2 – Exemple de patrons générés à partir du corpus

- La Détection de la Présence d'un Concept (DPC) : le plus souvent, ce sont des patrons constitués par des termes simples (patrons simples) utilisés pour signaler la présence d'un concept. Ce type de patron est appliqué sur les marqueurs dépourvus d'ambigüité.
- La Désambigüisation entre deux concepts (DEC) : ce type de patron est utilisé pour détecter le type d'entreprise qui cause la principale ambigüité dans l'analyse. Une mauvaise détection du type d'entreprise peut engendrer beau-

coup d'autres ambigüités (patrons enrichis). Ils permettent la classification de l'entreprise parmi l'une des deux principales classes conceptuelles (Réalisation d'outillage de production ou Fabrication des produits manufacturiers). Evidemment, certaines entreprises peuvent être classées dans les deux types puisqu'elles peuvent effectuer les deux types de productions. Ce type de patrons repose surtout sur les marqueurs production et ingénierie (verbe d'action) qui s'insèrent dans des patrons enrichis comme suit :
Représentant Entreprise - verbe d'action - COD
Ce patron nécessite l'extraction et l'analyse du COD dans la phrase puisque c'est lui qui va déterminer le type du produit délivré par l'entreprise et par conséquence le type de l'entreprise.

– L'Extraction d'une Information Complémentaire Rattachée au Concept (EICRC) : certains patrons sont utilisés pour extraire de l'information (patrons enrichis). Par exemple avec le marqueur *spécialiser* on cherche à extraire la spécialité de l'entreprise et non pas une simple détection de la présence d'une spécialité. C'est pourquoi le patron proposé est :
Représentant Entreprise - forme verbale passive incluant spécialisé - PREP-GN
Nous avons besoin d'extraire le GN pour savoir quelle est la spécialité de l'entreprise. Le résultat de l'acquisition des patrons à partir du corpus constitue une bibliothèque de 35 patrons enrichis et de 100 patrons simples non ambigus (voir annexe).

11.3 Transcodage des patrons

C'est l'écriture des patrons dans un langage formel compréhensible par la machine. Comme nous utilisons Unitex pour projeter des patrons sur le corpus, le transcodage est fait sous cet environnement. Ainsi les patrons sont décrits sous la forme de grammaires locales qui représentent un moyen puissant pour représenter la plupart des phénomènes linguistiques.
Unitex permet de représenter un ensemble d'expressions linguistiques sous forme d'un automate. Dans la représentation proposée, les graphes contiennent les éléments du vocabulaire dans des *boîtes* correspondant aux états de l'automate. Unitex per-

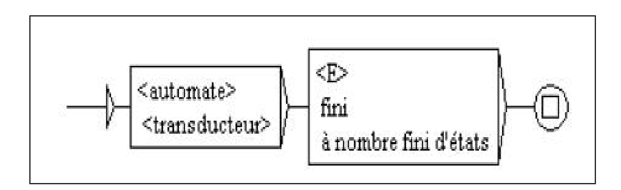


FIGURE 11.2 – Automate sous UNITEX

met également de modéliser des automates par des réseaux de transitions récursifs (RTN), où un état correspond en fait à un sous-ensemble appelé dynamiquement. L'appel à un sous-graphe apparait en grisé. Le graphe 11.3 suivant est équivalent

au graphe 11.2 s'il existe des automates appelés "automate" et "fini" équivalents. Les automates peuvent subir l'opération étoile, ainsi que l'union, l'intersection et le

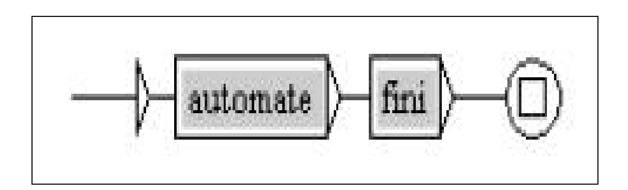


FIGURE 11.3 – Automates récursifs sous UNITEX

calcul complémentaire.
Dans notre contexte, nous nous dotons de deux types de patrons : patrons simples, composés généralement d'un seul terme ou d'une expression simple, et patrons enrichis composés des structures linguistiques plus au moins complexes. Ci-dessous des exemples de transcodage de ces deux types de patrons :

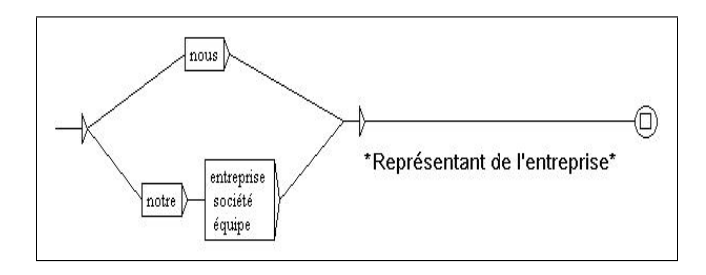


FIGURE 11.4 – Exemple de patron simple

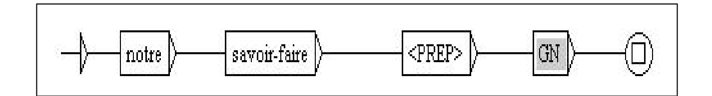


FIGURE 11.5 – Exemple de patron enrichi

Unitex ne permet pas de détecter un groupe nominal, c'est pourquoi nous avons créé un automate qui permet la reconnaissance du groupe nominal et le complément d'objet direct dans une phrase. A chaque fois que l'on a besoin d'un GN ou un COD dans un patron, on les appelle dynamiquement grâce à leurs sous-graphes.

11.4 Projection des patrons sur le corpus

La projection des patrons sur le corpus se fait par la recherche des occurences des schémas linguistiques, traduite sous la forme d'automates, dans le texte de l'entreprise. On se base sur le programme *locate* d'Unitex qui permet cette projection. Voici (figure 11.6) un exemple de projection d'un patron enrichi qui permet de typer l'entreprise selon sa production.

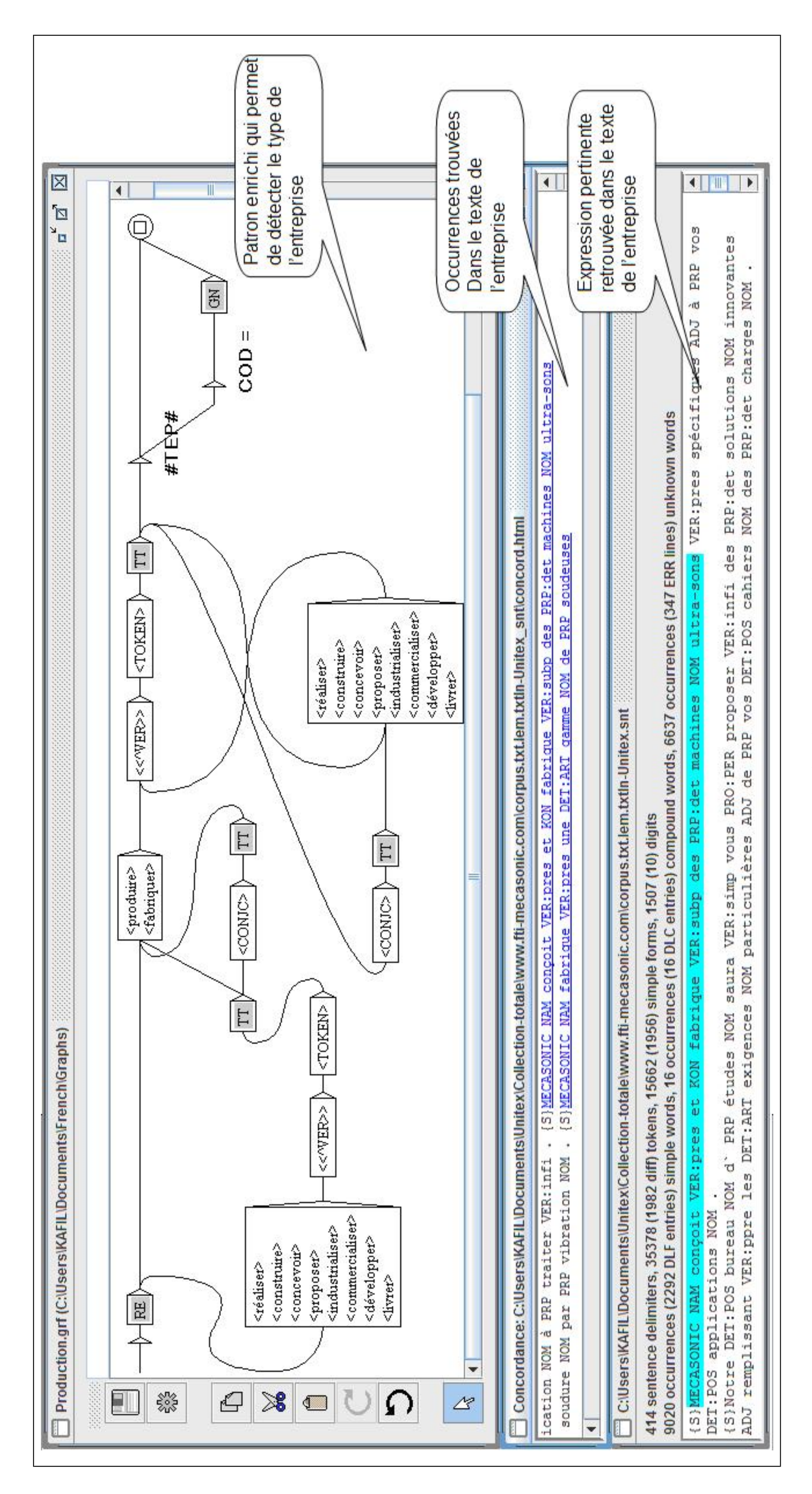


Figure 11.6 – Exemple de projection de patrons

Dans l'exemple de la figure 11.6, la projection du patron sur le texte de l'entreprise donne lieu à deux occurrences.
L'étape de projection des patrons sur le texte nous fournit une liste de patrons retrouvés pour chaque entreprise. Cette liste de patrons traduit un ensemble de concepts non ambigus qui a été détecté et retrouvé automatiquement dans le texte de l'entreprise. Chaque occurence trouvée par un patron est un élément constitutif qui vient s'ajouter pour construire la trace de compétence de l'entreprise.

Entreprise	Patrons retrouvés
www.mecadex.com	Analyse de la valeur Assemblage Atelier-de-production Bureau d'étude CAD CAO Caractéristiques techniques Certification Conception Décolletage Emboutissage Exigence Forge Fraisage Haute technologie Ingénierie concourante Montage Outils-moyens de mesure Partenariat Qualification Tournage Traitement de surface Usinage Spécialité
www.boisset-et-cie.fr	Appareils-équipements Cahier de charge CAO Conception Essais Habilité Haute précision Haute vitesse Ingénierie Outils-moyens de mesure Partenariat

Table 11.3 – Résultat de localisation des patrons dans quelques entreprises

La question qui reste à résoudre est comment traduire ces patrons en classes sémantiques de l'ontologie des traces de compétences.

11.5 Conclusion

Nous avons présenté la méthode suivie pour l'extraction des briques d'information pertinentes. Cette méthode est basée sur l'utilisation des patrons linguistiques pertinents vis-à-vis de notre recherche. La construction des patrons syntaxiques est faite autour des marqueurs qui représentent les instances de l'ontologie des traces de compétences par étude des corrélations entre ces derniers et les mots du corpus. La communication entre l'ontologie et le corpus est réalisée grâçe à ces patrons syntaxiques.
Dans la suite, nous allons détailler comment est réalisée la communication Ontologie-Ontologie (Métier-Générique) pour construire une trace complète des compétences de l'entreprise représentée sous la forme d'un sous-arbre de l'ontologie des traces de compétences des entreprises.

Chapitre 12

Performance du système d'extraction

12.1 Protocole d'Activation

L'étape d'activation consiste à transformer les patrons retrouvés en classes de concepts sémantiques de l'ontologie des traces des compétences. Il s'agit plus précisément d'activer les classes sémantiques de l'ontologie à partir de la présence ou non des patrons dans le texte de l'entreprise. La présence du patron traduit une détection d'un concept autour d'un marqueur. Pour réaliser cette activation, nous avons construit un algorithme basé sur des règles déterministes qui guident l'activation de la classe. Une classe est activée s'il y a au moins un patron détecté parmi la liste des patrons qui lui est attachée. Ainsi une classe-fils activée active la classe-père.

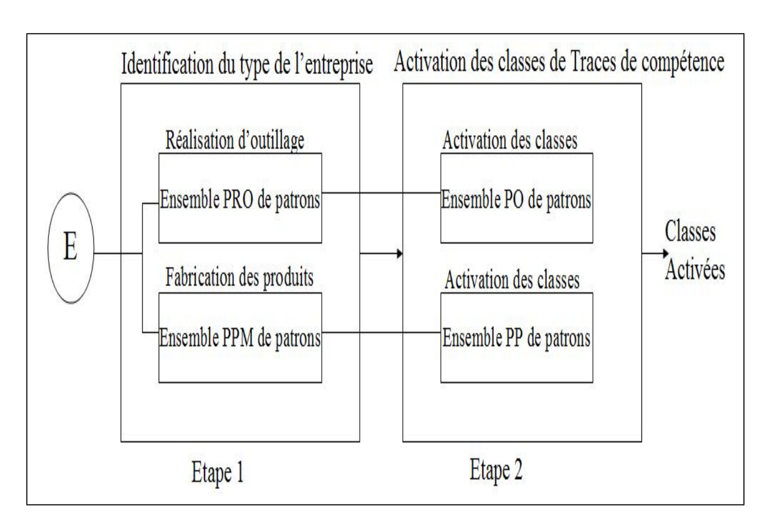


Figure 12.1 – Processus d'activation en deux étapes

Dans un premier temps, nous cherchons à détecter le type de l'entreprise. Ce type est retrouvé à partir d'un sous ensemble bien spécifique de patrons (Ex : Production, Ingénierie...). Cette première étape permet de classer l'entreprise parmi les deux classes conceptuelles de l'ontologie (Réalisation d'outillage de production ou Fabrication de produits manufacturiers). Nous avons déja signalé que certaines entreprises peuvent être classées dans les deux types puisqu'elles peuvent exercer les deux types de production. Par conséquence l'activation est faite par les deux types de patrons.

Dans un deuxième temps, selon le type détecté de l'entreprise, le processus d'activation se déroule autour d'un certain patron. Chaque patron est associé à chaque classe qu'il doit activer.
Ces deux étapes sont basées sur l'hypothèse qu'une entreprise, si elle réalise de l'outillage de production, ne peut pas avoir des traces de compétences autour de l'usage de procédés de production par exemple. Le type de produit qu'elle réalise est une information clé qu'il faut détecter et extraire en amont du processus d'activation, puisqu'elle permet de déterminer et de guider le chemin d'activation des classes des traces de compétences.
T ← Texte de l'entreprise
PT ← Liste des patrons permettant de typer l'entreprise
PRO ← Liste des patrons de type Production Réalisation d'Outillage
PPM ← Liste des patrons de type Production des Produits Manufacturiers
PO ← Liste des patrons à chercher pour Production Outillage
PP ← Liste des patrons à chercher pour Production Produits
ROP : "Réalisation d'Outillage de Production"
FPM : "Fabrication de Produit Manufacturiers"

```
Algorithm 1: PCA
begin
    foreach p in PT do
        if Recherche_Pattern(T, p) ⊂ PRO then
            Activer(ROP) = True;
            if Recherche_Pattern(T, p) ⊂ PPM then
                Activer(FPM) = True;
        else
            Activer(FPM) = True;
    if Activer(ROP) == True then
        foreach p in PO do
            if Recherche_Pattern(T, p) == True then
                Activer(Classe(p)) = True;
    if Activer(FPM) == True then
        foreach p in PP do
            if Recherche_Pattern(T, p) == True then
                Activer(Classe(p)) = True;
end
```

FIGURE 12.2 – Algorithme d'activation des classes ontologiques

Le but final du processus d'activation des classes est de fournir une trace des concepts traduite en un sous-arbre (relatif à chaque entreprise) de l'ontologie des traces des compétences. Avec deux sous-arbres de deux entreprises différentes, il devient possible de calculer une similarité.

12.2 Résultat de l'activation automatique

Le tableau 12.1 montre le résultat de l'activation automatique de deux entreprises en utilisant l'algorithme PCA (Patterns for Classes Activation).

12.3 Evaluation de l'activation

Pour évaluer les performances du système dans la phase d'activation, il faut se doter d'un ensemble d'entreprises pour lesquelles on connaît les compétences. En travaillant sur les données fournies au système (même ontologie, même texte), deux experts ont été chargés de lire le texte pour effectuer une activation manuelle basée sur la compréhension et l'interprétation du sens du texte vis-à-vis de l'ontologie du domaine. Cette tâche manuelle est fastidieuse et coûteuse en termes de temps, c'est pourquoi il a été choisi de faire une évaluation sur un sous-ensemble (10 entreprises) de la collection totale. Toutefois, si les résultats de l'activation automatique sont proches d'une telle activation manuelle faite par l'expert, nous aurons interêt à continuer l'activation automatique sur toute la collection. Cette décision dépendra des résultats de performance de la méthode d'activation [1].

1. on cherche bien une performance liée aux classes indépendamment du fait qu'elles soient activées par une ou plusieurs sous-classes ou patrons

Entreprise	Patrons retrouvés	Classes activées
www.mecadex.com	Analyse de la valeur	Production
	Assemblage	Usage de relation client
	Atelier-de-production	Usage de procédés d'assemblage
	Bureau d'étude	Usage de procédés fabrication
	CAD	Ingénierie
	CAO	Usage de procédés d'ingénierie
	Caractéristiques techniques	Usage de procédés CAO
	Certification	Caractéristiques techniques
	Conception	Ingénierie
	Décolletage	Décolletage
	Emboutissage	Emboutissage
	Exigence	Suivi des exigences
	Forge	Forgeage
	Forgeage	Forgeage
	Fraisage	Fraisage
	Haute technologie	Haute technologie
	Ingénierie concourante	Usage de PLM
	Montage	Montage
	Outils-moyens de mesure	Outillage de contrôle
	Partenariat	Partenariat
	Qualification	Qualification des produits
	Tournage	Tournage
	Traitement de surface	Traitement de surface
	Usinage	Usage de procédés d'usinage
	Usinage-laser	Laser
	Spécialité	Spécialisation
www.boisset-et-cie.fr	Appareils-équipements	Production
	Cahier de charge	Usage de procédés de fabrication
	CAO	Usage de relation client
	Conception	Usage de procédés CAO
	Essais	Ingénierie
	Habilité	Usage de procédés d'ingénierie
	Haute précision	Qualification des produits
	Haute vitesse	Habilitation
	Ingénierie	Caractéristiques techniques
	Outils-moyens de mesure	Outillage de contrôle
	Partenariat	Partenariat clé

TABLE 12.1 – Résultat de l'activation automatique des classes ontologiques pour deux entreprises

12.3.1 Activation des experts

Un premier travail d'expertise a été réalisé par deux experts. Lors de l'activation manuelle, chaque expert peut interpréter différemment les concepts ou simplement commettre des erreurs, ce qui fait que les experts ne sont pas forcément d'accord sur cette activation manuelle. L'activation considérée comme référence est le résultat d'un accord entre les deux experts après correction des erreurs. Avec cette activation de référence, nous allons évaluer l'activation du système et celle d'un expert seul pour dix entreprises dont le choix vérifie bien une diversité.

12.3.2 Evaluation de l'activation du système

L'évaluation de l'activation automatique est basée sur les deux indicateurs Précision et Rappel, en considérant que les classes qui sont activées par l'expert (après correction) sont les classes pertinentes qu'il faut activer pour chaque entreprise. Le tableau 12.2 résume la performance de l'activation automatique du système basée sur l'algorithme PCA :

Entreprise	**Précision**	**Rappel**
www.boisset-et-cie.fr	0.81	0.56
www.chambon.com	0.92	0.7
www.flip-elec.fr	0.87	0.5
www.martin-joseph.com	1	0.66
www.bargy-decolletage.com	0.75	0.54
www.entechno.fr	1	0.7
www.attax.com	0.76	0.83
www.fti-mecasonic.com	0.8	0.66
www.isojet.com	0.87	0.77
www.sic-marking.com	0.88	0.57
Moyenne	**0.87**	**0.64**

TABLE 12.2 – Evaluation de l'activation automatique

12.3.3 Evaluation de l'activation d'un expert

Entreprise	Précision	Rappel
www.boisset-et-cie.fr	0.93	0.87
www.chambon.com	0.73	0.64
www.flip-elec.fr	1	0.57
www.martin-joseph.com	0.72	0.88
www.bargy-decolletage.com	0.83	0.9
www.entechno.fr	0.75	0.6
www.attax.com	0.8	0.66
www.fti-mecasonic.com	0.88	0.84
www.isojet.com	1	0.77
www.sic-marking.com	0.91	0.78
Moyenne	**0.84**	**0.75**

TABLE 12.3 – Evaluation de l'activation de l'expert

Dans la table 12.3, la précision $0,84$ de l'expert provient du fait qu'il est inca-

pable d'effectuer une activation exhaustive (précision = 1) sur toutes les classes de l'ontologie métier dont le nombre est trop important. Le rappel est inferieur à 1 à cause de la tâche difficile de détection des marqueurs dans le texte (nombre important) et la mise en relation d'activation marqueurs-classes. À ces raisons s'ajoute le fait que deux experts peuvent avoir deux interprétations différentes sur les concepts du texte de l'entreprise, une activation dite "activation référence" a donc été créée pour corriger le désaccord entre les deux experts et converger vers une solution unique.

12.3.4 Synthèse d'évaluation de l'activation

Le tableau 12.4 décrit la comparaison entre l'activation automatique et l'activation manuelle montre bien que les résultats sont proches (0.84 et 0.87 ; 0 .75 et 0.64). Avec l'activation automatique on gagne une légère précision (0.87) car l'expert, lors du processus d'activation, ne peut pas faire une activation exhaustive sur toutes les classes : comme le nombre de classes à activer est important (50 classes) et les classes sont parfois sémantiquement très proches, il est possible que la même information que l'expert retient d'après sa compréhension du texte de l'entreprise n'active pas la même classe d'une entreprise à une autre. L'expert a fait un rappel de 0.75 dans son activation manuelle ; la manque de 0.25 de non activation des classes souligne bien la difficulté de cette tâche dûe au type de texte analysé (complexe, très hétérogène, mal structuré).

Activation	Précision	Rappel
Expert	0.84	0.75
Système	0.87	0.64

Table 12.4 – Synthèse de l'évaluation de l'activation automatique

On constate une légère diminution au niveau du Rappel avec l'activation automatique (l'activation de l'expert est basée sur une compréhension humaine du texte de l'entreprise). Par contre, l'activation du système utilise l'ensemble des marqueurs (instances) de l'ontologie des traces de compétences. Nous estimons qu'un enrichissement automatique ou semi-automatique des instances de l'ontologie pourraît être capable de pallier cet écart de différence au niveau du rappel.
Compte tenu de la tâche fastidieuse de l'activation manuelle des classes de l'ontologie des traces de compétences (c'est la raison pour laquelle l'évaluation est faite seulement sur dix entreprises). L'activation automatique est recommandée lorsqu'on augmente le nombre d'entreprises.

Partie 4 : Synthèse des Résultats

Cette dernière partie sort du cadre de la stricte recherche en informatique. Les résultats propres à la recherche sont discutés dans les précédents chapitres. Dans cette partie, nous discutons leur applicabilité dans un objectif de génie industriel qui nous a fourni le contexte applicatif. Ainsi, cette partie est une continuité des deux précédentes. Elle combine leurs résultats pour construire une cartographie théorique des modes de coordination au sein d'un réseau d'entreprises. Le résultat de la deuxième partie permet d'identifier les sous-groupes d'entreprises ayant des activités complémentaires. Le résultat de la troisième partie permet d'extraire des traces de compétences des entreprises. Ces deux résultats sont analysés dans cette partie pour identifier un mode de coordination préférentiel entre les entreprises du réseau.
Le premier chapitre de cette dernière partie explicite l'application des résultats trouvés par les deux systèmes d'extraction (SEI-1 et SEI-2) dans le contexte de la construction de réseaux d'entreprises. Nous explicitons la méthode utilisée pour quantifier l'éloignement entre les différentes traces de compétences afin de fournir l'information synthétique (similarité entre les compétences) indispensable à l'application de la méthode d'aide à la décision pour la construction des réseaux d'entreprises. Le deuxième chapitre présente nos conclusions et nos perspectives sur l'ensemble des contributions répondant aux objectifs de la thèse.

Chapitre 13

Application dans le contexte des réseaux d'entreprises

13.1 Introduction

L'analyse de la similarité des compétences au sein du réseau permet de préciser quelles sont les entreprises du réseau qui pourraient se coordonner dans une logique de réseau, et le type de mode de coordination pertinent, en fonction de la complémentarité de leurs activités. Cette analyse doit s'appuyer sur des outils mathématiques pertinents : comme justifié précedemment, nous utilisons des méthodes développées dans la thèse de M. Benali [15]

Nous nous intéressons ici à la modélisation du concept de similarité des compétences dans un réseau d'entreprises. L'objectif est d'appliquer une méthodologie qui nous permet d'isoler les sous-ensembles d'entreprises ayant des compétences proches. Pour modéliser et quantifier les compétences, nous utilisons les résultats de l'extraction du SEI-2 qui fournit des traces de compétences décrites en sous-arbres de l'ontologie des traces des compétences. Des notions de similarité basée sur la distance de Hamming sont utilisées pour quantifier l'éloignement entre les différentes traces des compétences des entreprises. Finalement, des outils d'analyse de données permettent d'identifier les sous-ensembles d'entreprises les plus proches en termes de compétences. Les expérimentations réalisées dans ce chapitre sont uniquement données à titre d'illustration et dans un but de test de faisabilité. Elles seront réalisées seulement sur un échantillon de 10 entreprises pour ne pas encombrer la cartographie finale du réseau.

13.2 Trace de Compétence d'une entreprise

Une trace de compétence d'une entreprise est un sous-arbre de l'ontologie des traces des compétences. Elle décrit l'ensemble des classes conceptuelles qui ont été activées par le résultat des informations extraites par SEI-2. Pour donner une idée claire sur la définition et la structure d'une trace de compétence, nous modélisons dans l'arbre ontologique de la figure 13.1 la structure globale de l'ontologie des traces des compétences pour les capacités techniques. Elle contient 4 niveaux dont chacun contient différentes classes conceptuelles. Chaque classe est représentée par un cercle. Par exemple, le premier niveau représente six classes ontologiques qui sont respectivement les suivantes (voir Annexe B pour l'ontologie complète) :

1. Traces du domaine technologique
2. Traces des ressources et des procédés de conception
3. Traces des ressources et des procédés de production
4. Traces de qualité et des performances des produits/services
5. Traces d'innovation sur les procédés techniques
6. Traces de démarche qualité sur les processus et l'organisation

Chacune de ces classes contient des sous-classes. Dans la figure 13.1, sont représentées toutes les classes et sous classes de l'ontologie des traces des compétences, des capacités techniques. Les marqueurs qui permettent d'activer les classes grâce à leur description dans des schémas syntaxiques (patrons) ne sont pas représentés. Les cercles en vert sur la structure de l'ontologie représentent une trace réelle de

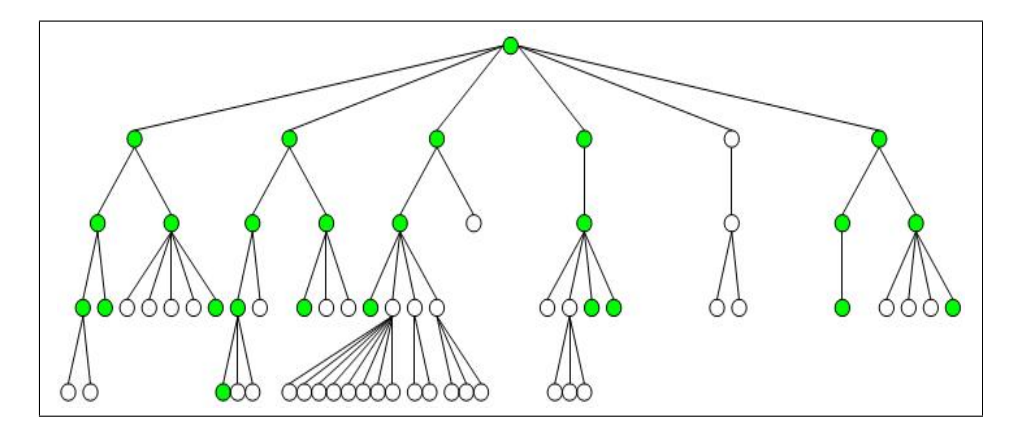


FIGURE 13.1 – Structure de l'ontologie sur le potentiel des capacités technique

compétence de l'entreprise Boisset [1], qui est du secteur de la mécanique. Les traces des compétences schématisées en vert illustrent bien les principes d'activation des classes ontologiques dont l'un d'entre eux induit que chaque sous-classe fille activée active son père.

13.3 Similarité des compétences entre deux entreprises

Dans notre travail, nous postulons que la similarité entre les traces des compétences d'entreprises extraites par le système UNICOMP évalue de manière relativement fiable la similarité des compétences réelles d'entreprises. Cette hypothèse est justifiée par le fait qu'une trace de compétence est construite autour des concepts de la compétence réelle d'une entreprise. L'ontologie des traces des compétences est construite à partir d'une modélisation de la compétence réelle dans une entreprise et à partir de l'extraction des concepts sur la notion de compétence réelle à partir d'un texte écrit par l'entreprise elle-même. Toutefois nous ne pouvons pas vérifier que les traces des compétences extraites sont les compétences réelles de

1. www.boisset-et-cie.fr

l'entreprise, puisque nous ne pouvons pas effectuer l'extraction d'une carte détaillée des compétences de l'entreprise à partir de son site web. L'aspect pragmatique dans l'extraction des traces des compétences, basé sur une analyse contextuelle du site web de l'entreprise, et les limites du corpus utilisé caractérisé par l'absence d'une description détaillée des compétences d'entreprises, nous amènent à identifier la similarité des traces des compétences avec la similarité des compétences des entreprises.

Comme nous l'avons exposé dans la section précédente, la trace de compétence d'une entreprise est décrite sous forme d'un sous-arbre de l'ontologie des traces de compétence. La similarité de deux entreprises en termes de compétences va donc être évaluée dans notre cas par une comparaison de deux sous-arbres ontologiques. Pour ce faire nous utilisons des calculs de distance, qui permettent de quantifier l'éloignement des traces de compétences entre deux entreprises différentes.

13.3.1 Mesure de similarité entre deux concepts ontologiques

Dans la littérature plusieurs travaux se sont intéressés à la mesure de similarité sémantique entre deux concepts d'une même ontologie. On peut distinguer trois grandes familles d'approches. Les approches basées sur les nœuds [136] [107] [86] utilisent des mesures de contenu informationnel pour détecter la similarité conceptuelle. La notion du contenu informationnel (CI) a été initialement introduite par [Res95] qui a montré qu'un mot est défini par le nombre des classes spécifiées et que la similarité sémantique entre deux concepts est quantifiée par la quantité d'information qu'ils partagent. La formule de Resnik est :

$$Sim(X,Y) = Max[E(CS(X,Y))] = Max[-log(P(CS(X,Y))])$$

où CS(X,Y) représente le concept le plus spécifique qui subsume les deux concepts X et Y. P est la probabilité de trouver une instance du concept c. La probabilité d'un concept c est calculée en divisant le nombre des instances de c par le nombre total des instances dans le corpus.

La deuxième famille d'approches repose sur la hiérarchie ou sur les distances des arcs [103] [52]. Le calcul de similarité utilise l'idée suivante : plus le chemin entre deux nœuds est court, plus ils sont semblables. Dans cette approche, les arcs représentent des distances uniformes, d'où que tous les liens sémantiques sont suposés possèder le même poids, ce qui rend délicate la définition des distances des liens.

Par exemple, le principe de calcul de similarité de Wu et Palmer [166] est basé sur les distances (N1 et N2) qui séparent les nœuds X et Y du nœud racine et la distance (N) qui sépare le concept subsumant (CS) [2] de X et de Y du nœud racine :

$$Sim(X,Y) = \frac{2N}{N1 + N2}$$

2. le concept commun le plus spécifique

La troisième famille d'approches est hybride [102] [137] et combine les deux premières.

$$Sim(X,Y) = -log(\frac{cd(X,Y)}{2M})$$

C'est la mesure de Leacock [102] où M est la longueur du chemin qui sépare le concept racine de l'ontologie du concept le plus bas. On note par cd (X, Y) la longueur du chemin le plus court qui sépare X de Y.

13.3.2 Similarité entre des sous-arbres ontologiques

La mesure de distance entre ontologies est utilisée dans l'espace ontologique pour décider comment mettre les ontologies en correspondance. De telles distances peuvent mesurer la facilité avec laquelle un alignement sera produit (sa rapidité et sa qualité). Le processus d'alignement d'ontologies a pour objectif de mettre en correspondance deux ontologies (ontologie source et ontologie cible). Une mise en correspondance (appelée aussi mapping) consiste à mettre en relation un concept de l'ontologie source avec un concept de l'ontologie cible pour obtenir une relation (is-a, part-of, etc.)
Nous ne sommes pas dans le cas d'un alignement entre ontologies, puisque nous ne cherchons pas à mettre en correspondance les concepts en les reliant par des relations. Nous cherchons à comparer deux sous-arbres d'une même ontologie. Cependant, toute mesure de distance conçue pour mettre les ontologies en correspondance peut être utilisée comme distance. Dans la suite, nous considérons quelques exemples de distances utilisées pour comparer deux ontologies.

13.3.2.1 Distances lexicales

Une distance entre deux ontologies peut être calculée à partir des étiquettes apparaissant dans les deux ontologies en utilisant une mesure telle que la distance de Hamming. Soient o et o' deux ontologies et $L()$ une fonction retournant les noms des entités dans une ontologie, la distance de Hamming sur les noms des classes est définie par :

$$D(o,o') = 1 - \frac{|L(o) \cap L(o')|}{|L(o) \cup L(o')|}$$

C'est une dissimilarité normalisée qui est relativement facile à calculer. Une extension de cette mesure consiste à utiliser des techniques de recherche d'information pour considérer tous les noms de l'ontologie comme une dimension. Chaque ontologie est prise comme un point dans un espace métrique de grande dimension, et une distance Euclidienne, ou cosinus, peut être calculée entre ces points. Des mesures de $tf - idf$ peuvent être utilisées pour évaluer la pertinence d'une ontologie vis-à-vis d'une autre.
Les mesures lexicales sont utilisables. Mais elles dépendent du langage utilisé. Si l'ontologie est exprimée en différents langages, cette mesure montre ses limites.

13.3.2.2 Mesures structurelles

La distance structurelle est fondée sur le calcul des distances entre les concepts [8] [162]. A partir d'une telle distance, nous pouvons définir une distance entre les ontologies. Parmi les mesures utilisées pour passer d'une distance entre les concepts à une distance entre les ontologies, on trouve :

- Distance de Hosdorff
- Lien moyen
- Distance de couplage maximal de poids minimal

13.3.2.3 Mesures sémantiques

La distance sémantique est fondée sur l'interprétation des ontologies. De telles mesures se fondent sur la notion de conséquence. Les mesures sémantiques ont été motivées par le traitement automatique du langage [51] a combiné l'utilisation d'un thésaurus créé automatiquement à partir d'un corpus textuel et Wordnet en utilisant une mesure de similarité sémantique pour trouver un sens prédominant des mots dans les textes non structurés. Les travaux de Hirst [82] ont étudié l'utilité des mesures sémantiques dans la correction automatique des erreurs d'orthographe.

13.3.3 Mesure utilisée

La mesure utilisée dans notre travail a comme objectif de détecter et d'évaluer l'éloignement entre deux sous arbres d'une même ontologie. Nous avons choisi de travailler avec une mesure simple basée sur la distance de Hamming. Ce choix est justifié par la difficulté de calculer le contenu informationnel d'un concept basé sur la probabilité de retrouver un concept, vu la qualité du corpus, ainsi que par la simplicité de mise en œuvre de cette mesure.
Cependant, nous avons enrichi la formule par un indicateur de profondeur (P) qui donne du poids à chacune des classes de l'ontologie selon sa profondeur. L'idée est de privilégier une intersection entre les deux ontologies (deux sous-arbres) puisque, dans le protocole d'activation, une classe fille active automatiquement sa classe père. La formule de similarité devient :

$$\delta(o,o') = 1 - \frac{\sum_{P=1}^{P=4} P\,|L_p(o) \cap L_p(o')|}{\sum_{P=1}^{P=4} P\,|L_p(o) \cup L_p(o')|} \tag{13.1}$$

La fonction δ est réelle positive et normalisée, c'est une dissimilarité qui est d'autant plus élevée que les ontologies diffèrent. Elle vérifie bien les propriétés suivantes :
$\forall o, o' \in O, \delta(o,o') \geq 0(positivit)$
$\forall o \in O, \delta(o,o) = 0(minimalit)$
$\forall o, o' \in O, \delta(o,o') = \delta(o',o)(symtrie)$
Cette mesure basée sur la distance de Hamming nous permetre de développer qualitativement la similarité entre les traces des compétences des différentes entreprises. Elle va transformer la similarité entre des classes conceptuelles des sous-arbres on-

tologiques en valeur normalisée dans l'intervalle $[0, 1]$. Plus les traces se ressemblent, plus cette valeur est proche de zéro.

13.4 Calcul de similarité pour un échantillon d'entreprises

La distance de Hamming relative présentée dans la section précédente est celle que nous utilisons pour quantifier l'éloignement des traces de compétences générées par le SEI-2. A l'aide d'un programme qui prend en entrée les traces des entreprises et l'ontologie globale des traces de compétences nous obtenons une matrice de valeurs (voir section 13.1). Nous avons construit cette matrice pour un ensemble constitué de dix entreprises sur lesquelles nous allons étudier le rapprochement en termes de compétences (table 13.1).

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
E1	0	0.68	0.5	0.7	0.63	0.71	0.69	0.62	0.52	0.64
E2	0.68	0	0.68	0.72	0.56	0.54	0.53	0.34	0.7	0.52
E3	0.5	0.68	0	0.64	0.62	0.62	0.56	0.6	0.62	0.5
E4	0.7	0.72	0.64	0	0.62	0.78	0.79	0.71	0.77	0.66
E5	0.63	0.56	0.62	0.62	0	0.67	0.56	0.59	0.55	0.41
E6	0.71	0.54	0.62	0.78	0.67	0	0.68	0.67	0.79	0.74
E7	0.69	0.53	0.56	0.79	0.56	0.68	0	0.45	0.65	0.46
E8	0.62	0.34	0.6	0.71	0.59	0.67	0.45	0	0.52	0.43
E9	0.52	0.7	0.62	0.77	0.55	0.79	0.65	0.52	0	0.48
E10	0.64	0.52	0.5	0.66	0.41	0.74	0.46	0.43	0.48	0

TABLE 13.1 – Mesure de similarité des compétences entre dix entreprises

Après avoir obtenu la matrice de distance, il est possible de classer les entreprises selon trois types d'intensité. Des coupes horizontales sur le graphe de similarité de compétence (figure 13.2) nous permettront d'obtenir des sous-groupes avec des similarités plus ou moins fortes. Plus la coupe horizontale se situe vers le bas, plus la similarité dans les sous-groupes est forte, et plus les liens entre les entreprises sont forts. Les coupes horizontales nous permettrons de graduer les liens de type "Réseau Réactif". Le graphe de similarité de compétence permet non seulement de définir le mode de coordination, mais aussi d'évaluer son intensité. Nous allons choisir 3 coupes, ce qui signifie que nous obtiendrons trois degrés d'intensité de la coopération de type réseau réactif : forte, moyenne et faible.

Première coupe (intensité forte) similarité $\leq$ 0,5 les paires d'entreprises sont :
{E1, E3}{E2, E8}{E3, E10}{E5, E10}{E7, E8}{E7, E10}{E8, E10}{E9, E10}

Deuxième coupe (intensité moyenne) 0,5$\prec$ similarité $\leq$0,7 les paires d'entreprises sont :
{E1, E2}{E1, E4}{E1, E5}{E1, E7}{E1, E8}{E1, E9}{E1, E10}{E2, E3} {E2,

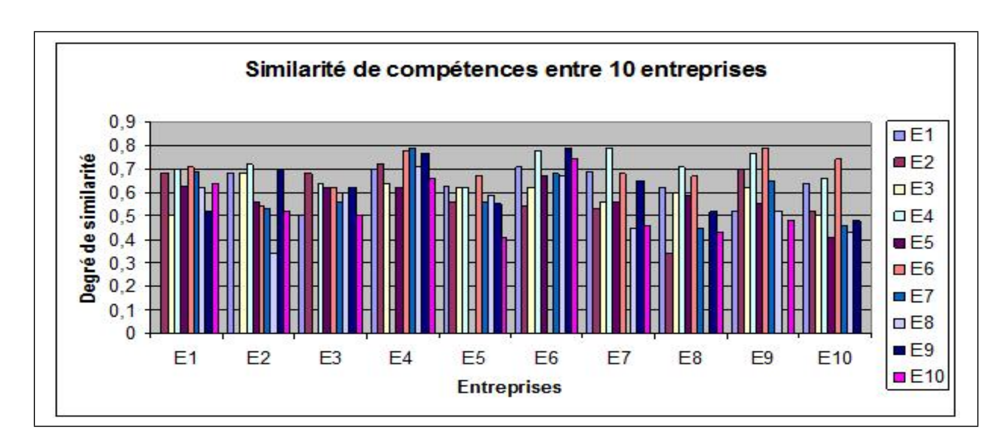


FIGURE 13.2 – Graphe de similarité de compétence

E5}{E2, E6}{E2, E7}{E2, E9}{E2, E10}{E3, E4}{E3, E5}{E3, E6}{E3, E7} {E3, E8}{E3, E9}{E4, E5}{E4, E10}{E5, E6}{E5, E7}{E5, E8}{E5, E9}{E6, E7}{E6, E8}{E7, E9}{E8, E9}

Troisième coupe (intensité faible) similarité $\succ$ 0,7 les paires d'entreprises sont : {E1, E6}{E2, E4}{E4, E6}{E4, E7}{E4, E8}{E4, E9}{E6, E9}{E6, E10}

La mesure de similarité choisie a un pouvoir de discrimination entre les différentes traces de compétences des entreprises. Les trois coupes horizontales permettent d'identifier les sous-groupes proches en termes de compétences avec des niveaux fort, moyen, et faible. Les résultats confirment aussi que les entreprises qui ont une similarité forte ne sont pas forcément du même secteur d'activité. De fait, deux entreprises du même secteur d'activités peuvent avoir des compétences différentes par exemple au niveau de l'utilisation des technologies du domaine.

13.5 Application de SEI-1 et SEI-2 pour la Construction des réseaux

Nous devons intégrer dans la représentation graphiques des réseaux d'entreprises le mode de coordination c'est-à-dire le type de relation qui peut exister entre deux entreprises et le niveau d'intensité de la relation. Pour cela nous allons commencer par définir les différents modes de coordination qui peuvent exister entre deux entreprises avant de commencer à construire le réseau.

13.5.1 Typologie des réseaux selon une analyse par activités et compétences

Dans sa thèse [15], Mehdi Benali a proposé une classification des réseaux que nous réutilisons dans le cadre de cette application : *"La complémentarité des activités et la similarité des compétences sont les deux facteurs clés pour l'analyse*

des modes de coordination industriels. Lorsque les activités sont complémentaires et les compétences sont similaires, le mode de coordination industriel le plus efficace semble être la direction hiérarchique au sein d'une firme. Par contre, quand des activités sont complémentaires et que les compétences sont non similaires, le mode de coordination le plus fréquent est la coopération inter-firmes (firme-réseau ou réseau de firmes). Ce type de coopération inter-firmes est nommé "Réseau Proactif" (RP), car les entreprises travaillent ensemble le long de la chaîne de valeur pour anticiper les besoins du marché et assurer une forte valeur ajoutée et souvent un haut degré d'innovation. Le deuxième type de coordination est assuré par des réseaux qualifiés de " Réseaux Réactif " (RR) qui correspond à des activités non complémentaires impliquant des compétences similaires [32]".

Ces réseaux réactifs sont souvent formés pour répondre à des motivations relatives à une réduction de coût par l'atteinte d'une taille optimale [130]. Ils ont comme objectif d'apporter une réponse collective aux contraintes et changements de l'environnement économique comme le partage des ressources, la centralisation de fonction... Le tableau suivant résume l'analyse exposée ci-dessus : Cette approche

	Activités non complémentaires	Activités complémentaires
Compétences non similaires	MARCHE	RESEAU PROACTIF
Compétences similaires	RESEAU REACTIF	FIRME

TABLE 13.2 – Typologie des réseaux selon une analyse par activités et compétences [15]

basée sur les activités et les compétences permet de construire un plan d'analyse des modes de coordination selon les axes Marché *vs* Firme et Réseaux Proactifs *vs* Réseaux Réactifs.

Cette typologie permet donc de préconiser des modes de coordination privilégiés entre deux entreprises d'un même réseau. Ainsi, deux entreprises ayant des activités complémentaires et des compétences similaires auront intérêt à mettre en place entre elles des liens de type Réseau Proactif.

13.5.2 Illustration de la construction des réseaux

13.5.2.1 Méthode de construction Données-Information-Connaissances

Le schéma suivant résume les étapes suivies dans la démarche de construction d'une cartographie des réseaux d'entreprises à partir de leurs sites web. Cette méthode est basée sur un modèle de transformation des données (sites web des entreprises) en informations (complémentarité des activités et trace de compétence) et des informations en connaissances (des réseaux d'entreprises en coopération) :

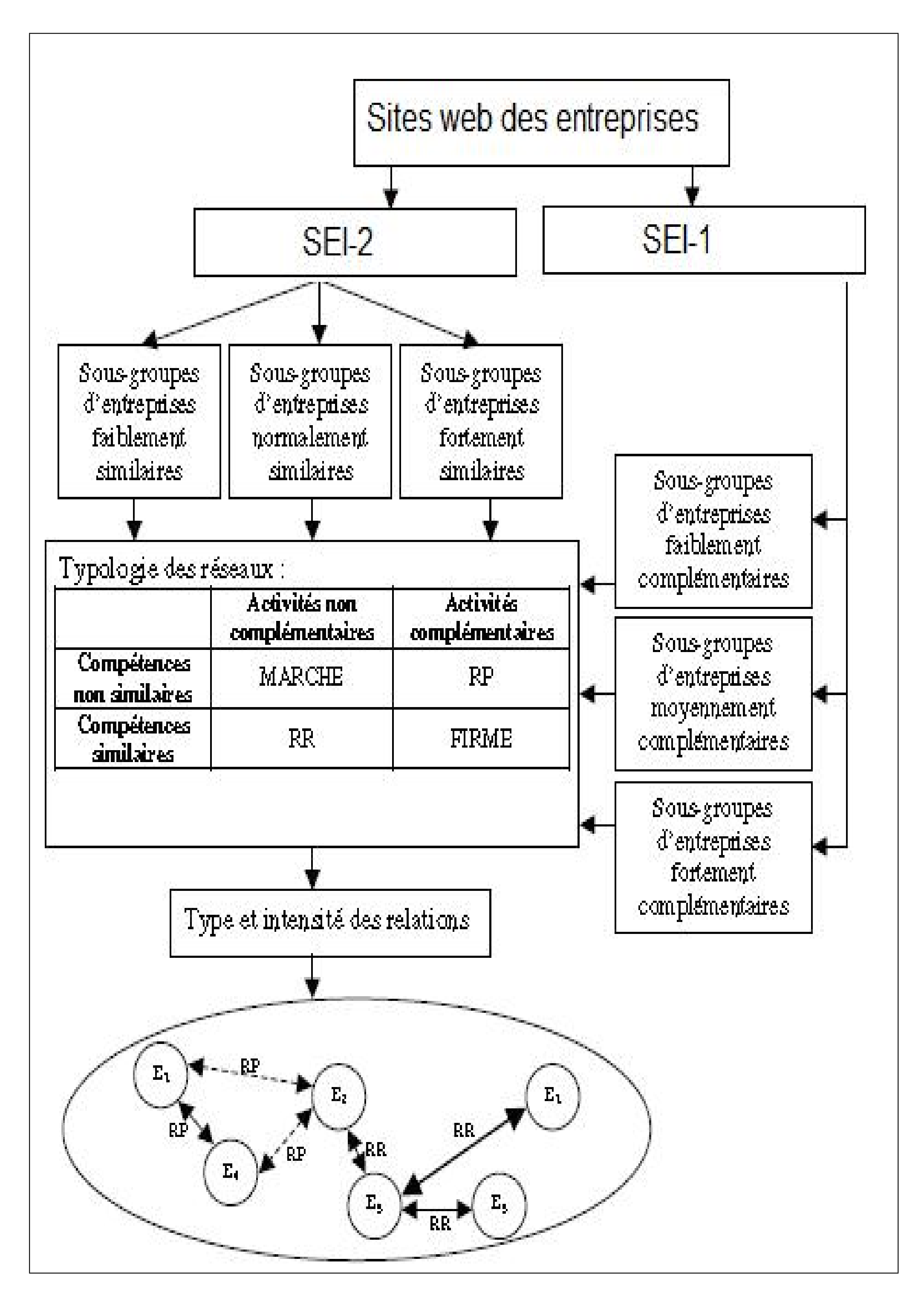


FIGURE 13.3 – Schéma de la méthodologie pour la construction d'une cartographie adaptée des travaux de Benali [15]

13.5.2.2 Exemple d'application

Nous appliquons, dans cette section, la méthodologie complète de construction de la cartographie du réseau de 10 entreprises.

Analyse de la complémentarité des activités

L'analyse de la complémentarité des activités pour les 10 entreprises a fait ressortir les résultats suivants (L'application de cette méthode de partitionnement est présentée dans le chapitre 7. Parmi les 25 entreprises nous ne considérons que 10 entreprises).

5	0.5	2	{E9, E8, E10} ;{E1, E3, E2, E4, E5, E7, E6}	0.44	Forte

Table 13.3 – Résultat du partitionnement pour la détection des entreprises en complémentarité d'activité.

Analyse de la similarité des compétences

L'analyse de la compétence sur les 10 entreprises fait resortir les résultats suivants Première coupe (intensité forte) similarité $\leq$ 0,5 les paires d'entreprises sont : {E1, E3}{E2, E8}{E3, E10}{E5, E10}{E7, E8}{E7, E10}{E8, E10}{E9, E10}

Construction de la cartographie

Nous allons maintenant construire la cartographie du réseau à partir des résultats obtenus ci-dessus et de la typologie des réseaux proposée. Dans la construction, nous n'allons pas prendre en compte les liens de faible et de moyenne intensité pour ne pas encombrer le graphique. La méthode de construction consiste à mettre en place les liens de type Réseau Proactif (RP) et Réseau Réactif (RR) c'est-à-dire placer les liens entre les paires d'entreprises qui ont respectivement des activités complémentaires et des compétences non similaires et ainsi des activités non complémentaires et des compétences similaires. La carte obtenue est donnée dans la figure 13.4. Nous remarquons que le nombre de liens est élevé, cela est dû à la qualité de la coupe (coupe à un niveau fort). En effet, l'algorithme de partitionnement des activités complémentaires n'élimine pas les arcs avec un haut degré, ce qui a pour effet de garder les fortes complémentarités qui ainsi apparaissent dans la cartographie aux travers des liens de type Réseau Proactif forts. Cette caractéristique a aussi pour effet de faire ressortir les fortes synergies au travers de possibilités de fusion/acquisition. La représentation graphique a bien entendu ses limites. Elle devient illisible au bout d'un certain nombre d'entreprises et d'interconnexions. De plus, elle est exploitable seulement à l'oeil nu, et d'un point de vue mathématique il est préférable de recourir à une représentation matricielle. La représentation matricielle est utile pour repérer et évaluer le rôle et la position des noeuds par exemple. Elle nous aide à reconstituer la structure du réseau afin de le décomposer en blocs homogènes ou clusters. Il sera aussi possible de faire des comparaisons plus approfondies entre deux cartographies.

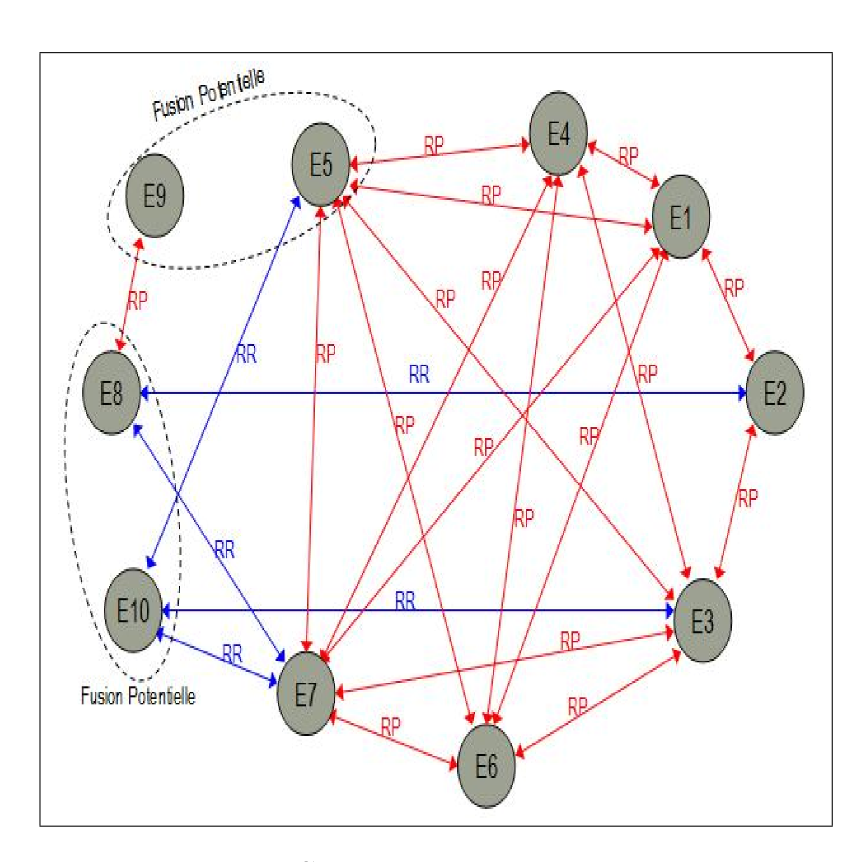


FIGURE 13.4 – Cartographie des réseaux d'entreprises

13.5.2.3 Analyse et utilisation de la cartographie

La cartographie organisationnelle obtenue à l'issue de notre méthode peut être utilisée par différents utilisateurs et pour diverses analyses.

1. *Différents utilisateurs* : une cartographie peut être utilisée par un utilisateur externe au réseau, une institution, un consultant ou tout simplement une entreprise qui veut intégrer le réseau.
   – Une institution, à travers la cartographie, peut avoir une idée du tissu des réseaux d'entreprises existants sur une région ou pour une filière industrielle, ce qui permet de construire une politique ou une stratégie mieux ciblée.
   – Un consultant peut analyser la structure des modes de coordination internes à un réseau pour améliorer les synergies, pour développer et améliorer les collaborations internes au réseau, pour trouver des potentialités de fusion/acquisition, pour mieux les contrôler.
   – Une entreprise qui veut intégrer un réseau peut avoir une idée de sa position future dans le réseau : elle peut détecter les entreprises avec lesquelles elle développera des coopérations.
2. *Types d'analyses* : plusieurs types d'analyses sont possibles. Nous pouvons citer :
   – Identification des potentialités de coopérations (opportunités futures entre les entreprises appartenant au réseau) à développer et qui ont un effet positif sur la performance des entreprises.
   – Identification des risques éventuels de fusion/acquisition et des coopérations qui peuvent échouer.
   – Préconisations en termes d'éléments influant la performance du réseau, comme par exemple des orientations pour structurer le système d'information et de communication entre les différents partenaires au sein du réseau.
   – Détection de sous-réseaux, des clusters au sein du réseau où la collaboration, les échanges, et les synergies sont potentiellement plus utiles.
   – Détection des entreprises qui jouent un rôle important dans le réseau (entreprises pivots).

Chapitre 14

Conclusion et Perspectives

14.1 Conclusion générale

Le point de départ de notre travail de thèse était un problème formulé dans un contexte de collaborations inter-entreprises, qui porte sur le traitement automatique de l'information pour la génération des connaissances. Tout au long de ce travail, nous avons présenté un ensemble d'observations, d'hypothèses, de réalisations et d'évaluations. La question qui s'impose naturellement est de savoir si nous avons répondu au problème de départ, c'est-à-dire l'élaboration d'une méthode et d'un système de recherche et d'extraction d'informations à partir du web (site web des entreprises) pour un objectif d'aide à la décision dans la construction de réseaux d'entreprises en collaboration. Nous sommes tentés de répondre par l'affirmative.

Les travaux présentés dans ce mémoire ont visé à confronter les techniques de traitement de l'information à la problématique de construction de réseaux d'entreprises en collaboration, en particulier par la recherche et l'extraction d'information à partir du web. La recherche d'information que nous effectuons se fait dans un environnement ouvert où les organisations ne se connaissent pas et ont une information hétérogène publique et non restreinte. Des travaux antérieurs au sein de notre laboratoire ont proposé une typologie des modes de coordination entre les différentes entreprises d'un réseau. Cette typologie est basée sur deux paramètres : la complémentarité des activités et la similarité des compétences. Ces deux paramètres ont été identifiés comme étant discriminants pour justifier le choix d'un type de coopération industrielle. C'est pourquoi notre besoin d'information s'articule autour de deux systèmes d'extraction d'information.

L'enjeu scientifique de la thèse est de contribuer à une automatisation de la recherche d'informations caractérisant des entreprises, en vue d'appliquer les modèles formels d'aide à la décision qui visent à identifier des collaborations inter-entreprises. Ainsi, l'objectif est d'expliciter la capacité à utiliser des ressources sémantiques propres au métier pour améliorer les performances des mécanismes de recherche et d'extraction d'information avec deux cas traités :

- Ressources sémantiques structurées disponibles propres au métier (SEI-1)
- Ressources sémantiques structurées non disponibles propres au métier (SEI-2)

Nous avons présenté une première approche basée sur des outils et des méthodes de recherche d'information, à savoir, l'indexation contrôlée et la mesure de similarité. Cette approche de SRI est mise en place pour la détection automatique du

secteur d'activité de l'entreprise à partir de son site web. Nous avons utilisé le code NAF comme un thésaurus qui reflète une représentation sémantique et conceptuelle de tous les domaines d'activités pour proposer un premier système d'extraction d'information qui permet la détection du domaine d'activité de l'entreprise à partir de son site web. La bonne connaissance du secteur d'activité permet de faire émerger des réseaux coopératifs d'entreprises de divers types.

Dans un deuxième volet, une deuxième approche a proposé le système UNICOMP, qui est dédié à l'extraction des traces de compétences des entreprises à partir de leur site web. Il prend en entrée le site web de l'entreprise et une ontologie générale décrivant toutes les compétences des entreprises. UNICOMP mobilise des techniques d'extraction puissantes utilisant principalement une ontologie du domaine, une bibliothèque de patrons qui décrivent des schémas syntaxiques de l'information pertinente liée au concept de compétence et un ensemble de programmes de traitement automatique de textes afin d'extraire une information dépourvue de toute ambigüité.

Des résultats expérimentaux ont été obtenus pour chacune de ces étapes. Une application des informations extraites sur les activités et les compétences a servi à la construction de réseaux d'entreprises en collaboration.

14.2 Perspectives

La question de la construction deentreprises en collaboration avec d'autres partenaires constitue un enjeu majeur de la survie des acteurs industriels. Les problématiques posées par la construction de l'entreprise virtuelle se présentent à différents niveaux et ceci, depuis l'identification du besoin d'information commun jusqu'au déploiement. Nous souhaitons dans ce chapitre revenir sur les deux grandes partie de notre contribution (détection des activités des entreprises et extraction des compétences) pour détailler nos perspectives sur chacune des parties et/ou discuter des possibilités d'amélioration des méthodes et outils proposés.

14.2.1 La détection des Activités

De nombreuses améliorations et perspectives peuvent être apportées aux techniques de la recherche et de l'extraction d'information décrites dans cette partie :

- Il serait nécessaire de faire un passage à l'échelle pour prendre en compte toutes les classes du code NAF, c'est-à-dire tester tous les domaines d'activités des entreprises. En parallèle, il faudrait tester la robustesse du système, évaluer les temps et la qualité des réponses. Ce passage à l'échelle offre une indépendance du domaine d'activité des entreprises analysées. En même temps, il nous faudrait faire un passage à l'échelle pour le nombre de requêtes. Dans notre travail nous nous sommes limités à une collection d'une centaine d'entreprises.

Il est important d'augmenter le nombre d'entreprises analysées et de diversifier leurs activités pour généraliser la méthode de recherche d'information.
- Il faudrait introduire un modèle de recherche d'information basé sur la reformulation des requêtes traitant le domaine d'activité. Les requêtes sont construites autour du mot "activité" qui est corrélé à chaque fois avec des termes jugés pertinents : dans un premier temps nous projettons la requête construite par le mot "activité", nous déterminons le terme le plus corrélé avec ce mot parmi les réponses fournies, puis nous réinjectons une nouvelle requête. L'objectif de cette proposition est de comparer les performances du système entre une recherche basée sur une ressource sémantique structurée du domaine et une expansion de la requête en utilisant le web. Cette expansion de requête est vue comme un traitement pour élargir le champ de recherche pour cette requête. Une requête étendue va contenir davantage de termes reliés.
- Les recherches concernant la détection et l'extraction du texte à partir des séquences vidéo et des images sont encore confrontées à de sérieux problèmes. Le problème principal peut être expliqué par la différence entre l'information présente dans un document et celle donnée par une séquence vidéo [1], ainsi que les méthodes de stockage de chaque type. Les images des séquences vidéo contiennent de l'information plus difficile à traiter. Le texte n'est pas séparé du fond. Il est soit superposé comme les sous-titres, soit inclus dans la scène de l'image. Le fond de l'image peut être très complexe ce qui empêche une séparation facile des caractères. De plus, contrairement aux documents écrits, les séquences vidéo contiennent de l'information très riche en couleurs. Enfin, le texte n'est pas structuré en lignes. Souvent quelques mots courts et déconnectés flottent dans l'image. Généralement le but d'inclure le texte dans une animation ou une séquence vidéo est de mettre en exergue cette information pertinente qui décrit une caractéristique importante de l'entreprise. Il est important de proposer une méthode d'extraction de texte dans des pages html à partir des vidéos, des images...

14.2.2 L'extraction des compétences

La problématique à laquelle nous nous sommes attaqués n'en est qu'à ses débuts. Chaque étape, chaque semaine de notre travail nous a ouvert un nombre considérable de perspectives que la communauté se doit d'explorer.
- Enrichir automatiquement l'ontologie des traces de compétences : proposer une approche basée sur le traitement automatique du corpus et d'une liste de concepts décrivant une première version de l'ontologie initiale à enrichir. Cette approche peut être basée sur les étapes suivantes : commencer par générer des règles d'association pour la détection d'une corrélation entre les concepts de l'ontologie et les mots du corpus. Puis enrichir automatiquement l'ontologie initiale par les concepts appris selon des paramètres validés expérimentalement. Une deuxième perspective consiste à étendre la méthode de création

1. ou équivalente : images, animation flash, etc...

de corpus d'apprentissage en générant des requêtes spécifiques lancées sur le web selon des critères propres au domaine traité. Parallèlement, il serait très intéressant d'observer si l'ordre entre les mots peut avoir un impact sur la corrélation. Traditionnellement, les approches de fouille de textes considèrent plutôt les n-grammes pour prendre en compte l'ordre entre les mots ou les caractères. L'avantage des n-grammes est bien entendu de retrouver des mots très proches (i.e. en fonction de la valeur de n). Le défaut de ces approches dans le contexte de fouille de données est qu'elles nécessitent que les mots soient très proches afin de les repérer. Notre idée est d'étendre l'approche en utilisant la notion de motifs séquentiels et permettre ainsi d'extraire des concepts qui sont proches sans être consécutifs.
- Evaluer la robustesse de la communication entre l'ontologie métier et l'ontologie générique en testant avec d'autres domaines d'activités (l'informatique par exemple).
- Dans notre travail, les patrons syntaxiques sont construits manuellement à partir des marqueurs qui présentent les instances de l'ontologie métier. Il serait très intéressant de pouvoir construire une méthode automatique permettant d'extraire un patron autour du concept. Une des idées qui peut être étudiée dans ce cadre est d'effectuer une analyse linguistique très fine sur la phrase ou l'ensemble des mots corrélés au concept pour détecter les éléments focus d'information.
- Une des perspectives techniques est de développer un outil industriel destiné à la construction de la matrice du mode de coordination et de la construction de réseaux. Cet outil peut contenir un module de communication directe avec les entreprises pour vérifier les résultats trouvés.

Dans notre travail, nous nous sommes limités à l'analyse des deux paramètres principaux pour le choix d'un mode de coordination (les activités et les compétences) cependant, il reste des paramètres secondaires qui ont leur influence sur l'émergence des relations de coopération (degré d'internationalisation, degré de diversification, taille de l'entreprise etc.).

Glossaire

Lemmatisation : opération consistant à extraire la forme canonique d'un mot (son lemme), ainsi qu'éventuellement d'autres informations morphologiques. Exemple : "entreprises" est au féminin pluriel, et a pour lemme entreprise.

Poids : importance d'un mot dans un énoncé ou dans un document.

Précision : taux de documents pertinents parmi tous les documents retrouvés par le système.

Rappel : taux de documents pertinents retrouvés par le système parmi l'ensemble des documents pertinents de la collection.

Sémantique : étude de la signification des énoncés, indépendamment de tout contexte.

Syntaxe : partie de la grammaire décrivant les règles par lesquelles se combinent en phrases les unités significatives (mots).

Pragmatique : étude de la signification des énoncés en lien avec le contexte (interlocuteurs, phrases précédentes, connaissance commune du monde,...).

Document : un document est un volume d'information auto-explicative

Index : un index est une représentation synthétique du contenu sémantique d'un document.

Indexation : L'indexation est le processus responsable de l'extraction du contenu sémantique d'un document et de la représentation de ce contenu sous la forme d'un index.

APE Activité Principale Exercée

DPC Détection de Présence d'un Concept

DEC Désambiguisation Entre Concepts

EICRC Extraction d'Information Complémentaire Rattachée au Concept

DRT Discourse Representation Theory

EI Extraction d'Information

ERP Entreprise Ressource Planning

GCA Graphe de complémentarité des activités

GN Groupe Nominal
IC Ingénierie des Connaissances
idf inverse document frequency

KDD Knowledge discovery in databases

LSA Latent Semantic Analysis

MLP Multi Layer Perceptron

NAF Nomenclature des Activités Françaises

NTIC Nouvelles Technologies de l'Information et de la Communication

RD Recherche Documentaire

RTO Ressources Termino Ontologiques
RI Recherche d'Information

SI Système d'Iformation

SVM Support Vector Machine

TAL Traitement Automatique de la Langue

tf term frequency

OV Organisation Virtuelle

QR Question Réponse

VCH Vocabulaire Contrôlé Hiérarchisé

VC Vocabulaire Contrôlé

VBE Virtual Breeding Environment

Annexe A

L'Ontologie Générique

Compétences Entreprise

- Capacités
  - Capacités Technologiques
    - Traces du domaine technologique
    - Traces de ressources et produits techniques de l'entreprise
      - Traces de ressources et procédés de conception
      - Traces de ressources de procédés de production
      - Traces de qualité et de performances des produits-services délivrés
    - Traces d'innovation techniques sur les procédés de l'entreprise
      - Traces d'innovation sur les procédés techniques
      - Traces de démarches qualité sur les processus

Annexe B

L'Ontologie Métier

Traces du domaine technologique

- Domaine technique
  - Production {production, Fabrication}
    - Réalisation d'outillage de production {production, fabrication}
    - Fabrication de produits manufacturiers {production, fabrication}
  - Ingénierie {ingénierie, conception, bureau d'étude}
- Spécialisation
  - Haute technologie {haute technologie, technologie de pointes, technologie innovante}
  - Spécialisation {Spécialisé, expérience, spécificité, savoir-faire}
  - Domaine d'application {domaine, production, conception}
  - Secteur clientèle {clientèle, clients, domaine, industries}
  - Partenaires clé {partenariat}

Traces des ressources et procédés de conception

- Conception mécanique
  - Usage de Procédés d'ingénierie {ingénierie, conception, bureau d'étude, bureau des méthodes}
    - Usage de procédés CAO {CAO, CAD}
    - Usage de procédés de prototypage {prototypage, prototype}
    - Usage de procédés maquettage virtuel {simulation, maquette virtuelle}
  - Usage de Procédés de mécatronique {mécatronique}
- Outils de gestion
  - Usage de Relation client {Relation client, Cahier des charges, analyse du besoin, Analyse de la valeur}
  - Usage de PLM {PLM, support logistique, gestion intégrée, Cycle de vie, Ingénierie simultanée, Ingénierie concourante}
  - Gestion de projet {Gestion de projets, Suivi de projets, accompagnement clients, gestion des délais, réactivité, écoute client}

Traces de ressources et procédés de production

- Usage de Procédés de production
  - Usage de Procédés de fabrication {Atelier, (atelier-unités-ligne-technologie) de production, Parc machine, Machine outils, Commande numérique, (Appareils - équipements) de production, Robots, Fabrication-Fabriquant}
  - Usage de Procédés d'usinage {Usinage, procédés d'usinage, Centre d'usinage}
    - Laser {Laser, Usinage Laser, Laser haute vitesse, Découpe Laser}

- Fraisage {fraiseuse, fraisage}
- Tournage {Tournage}
- Rectification {Rectifiage, Rectifieuse}
- Emboutissage {Emboutissage}
- Décolletage {Décolletage}
- Forgeage {Forgeage, forge}
- Chaudronnerie {Chaudronnerie}
- Usage de Procédés Assemblage {Assemblage, Intégrateur}
  - Soudage {soudage, soudeuse}
  - Montage {montage}
- Usage de Procédés Traitement de surface {Traitement de surface, Traitement des métaux}
  - Traitement thermique {Traitement thermique}
  - Revêtement de surface {Revêtement de surface, Peinture, Chromage, revêtement de métaux, revêtement métallique, revêtement céramique, protection de surface}
  - Traitement chimique {Traitement chimique}
- Gestion de production {GPAO, ERP, Gestion de production, gestion de flux, optimisation de production, optimisation des flux, gestion logistique}

Traces qualité et de performance des produits/services délivrés

- Signes de performance
  - Satisfaction {Satisfaction, Confiance, reconnaissance}
  - Fiabilité des produits/services {Fiabilité}
    - Suivi des exigences {exigence}
    - Maîtrise des délais {délai}
    - Qualité des produits/services {qualité, performance}
  - Qualification des produits {Qualification, qualifié, agréés, essais, tests}
  - Caractéristiques techniques {Haute précision, Haute vitesse}

Traces d'innovation sur les procédés techniques

- Réalisation d'innovation sur les procédés
  - Innover {innovation sur les procédés ou processus, optimisation}
  - Investir {investissement sur les procédés ou processus, acquisition de procédés, amélioration-optimisation des procédés processus}

Traces de démarche qualité sur les processus

- Usage de Ressources de contrôle qualité
  - Outillage de contrôle {(Moyens-outil-outillage-technologie) de mesure, (Moyens-outil-outillage-technologie) de contrôle, Contrôle qualité}
- Utilisation de démarche qualité
  - Démarche qualité {(Maîtrise, gestion) de la qualité, exigence qualité, Assurance qualité}
  - Certification {certification, certifié}
  - Norme {norme-normalisé, ISO}
  - Habilitation {habilité, Agréé, Agréments}

Annexe C

Bibliothèque de patrons

PATRONS	USAGE	CLASSE A ACTIVER
Représentant Entreprise - produire	Détection de la présence d'un Concept (DPC)	PRODUCTION - PROCEDES DE FABRICATION
Représentant Entreprise - fabrique	DPC	PRODUCTION - PROCEDES DE FABRICATION
Représentant Entreprise - verbe d'action - COD Désambigüisation Entre Concepts	(DEC)	REALISATION DOUTILLAGE DE PRODUCTION
Représentant Entreprise - verbe d'action - COD	DEC	FABRICATION DE PRODUITS MANUFACTURIERS
Ingénierie	DPC	INGENIERIE - PROCEDES D INGENIERIE
Conception	DPC	
Bureau d'étude	DPC	
Haute technologie	DPC	HAUTE TECHNOLOGIE
Technologie de pointes	DPC	
Technologie innovante	DPC	
Représentant Entreprise - forme verbale passive incluant spécialisé - PREP-GN	Extraction d'Information Complémentaire Rattachée au Concept (EICRC)	SPECIALISATION - DOMAINE D APPLICATION
Expérience - PREP - GN	EICRC	
DET :POS relatif à l'entreprise - spécificité - verbe d'autoréférence - GN	EICRC	
DET :POS relatif à l'entreprise - Savoir faire - [] - PREP - GN	EICRC	
DET :POS - Domaine - COD	EICRC	DOMAINE D APPLICATION
DET :POS - Domaine - Verbe d'autoréférence - COD	EICRC	
Représentant Entreprise - Production/conception - COD	EICRC	
Clientèle - GN	EICRC	SECTEUR CLIENTELE
Det - Domaine - verbe Etre - GN	EICRC	
Dans - Det - Domaine - de - GN	EICRC	

PATRONS	USAGE	CLASSE A ACTIVER
Simulation	DPC	USAGE DE PROCEDES MAQUETTAGE VIRTUEL
marquette virtuelle	DPC	
Mécatronique	DPC	PROCEDES DE MECATRONIQUE
relation client	DPC	RELATION CLIENT
relation - avec - nos- clients	DPC	
relation - avec - les- clients	DPC	
cahier de charge	DPC	
Analyse du besoin	DPC	
Analyse de la valeur	DPC	
PLM	DPC	PLM
support logistique	DPC	
gestion intégrée	DPC	
cycle de vie	DPC	
Ingénierie simultanée	DPC	
Ingénierie concourante	DPC	
gestion de projet	DPC	GESTION DE PROJET
suivi de projet	DPC	
Accompagnement clients	DPC	
Gestion des délais	DPC	
Réactivité	DPC	
Ecoute client	DPC	
Atelier de production	DPC	PROCEDES DE FABRICATION
unité-ligne-technologie de production	DPC	
DET :POS - parc de machine	DEC	
machine-outils	DPC	
commande numérique	DPC	
DET :POS - appareils-equipements	DPC	
appareils-equipements - de - production	DPC	
robot - de - soudure	DEC	
usinage	DPC	POCEDES D USINAGE
procédés d'usinage	DPC	
centre d'usinage	DPC	
usinage - laser	DPC	LASER
laser haute vitesse	DPC	
découpe laser	DPC	
fraisage	DPC	FRAISAGE
fraiseuse	DPC	
tournage	DPC	TOURNAGE
rectifiage	DPC	RECTIFICATION
rectifieuse	DPC	
emboutissage	DPC	EMBOUTISSAGE
décolletage	DPC	DECOLLETAGE
forgeage	DPC	FORGEAGE
forge	DPC	
chaudrennerie	DPC	CHAUDRONNERIE
Assemblage	DPC	PROCEDES ASSEMBLAGE
Intégrateur	DPC	

PATRONS	USAGE	CLASSE A ACTIVER
Soudage	DPC	SOUDAGE
Soudeuse	DPC	
montage	DPC	MONTAGE
Traitement de surface	DPC	POCEDES TRAITEMENT DE SURFACE
Traitement de métaux	DPC	
Traitement thermique	DPC	TRAITEMENT THERMIQUE
Revêtement de surface	DPC	REVETEMENT DE SURFACE
Peinture	DPC	
Chromage	DPC	
Revêtement de métaux	DPC	
Revêtement métallique	DPC	
Revêtement céramique	DPC	
Protection de surface	DPC	
Traitement chimique	DPC	TRAITEMENT CHIMIQUE
GPAO	DPC	GESTION DE PRODUCTION
ERP	DPC	
Gestion de production	DPC	
Gestion de flux	DPC	
Optimisation de production	DPC	
Optimisation des flux	DPC	
Gestion logistique	DPC	
satisfaction -[DET]- clients	DEC	SATISFACTION
satisfaction -PREP - DET :POS - clients	DEC	
satisfaction -ADJ - clients	DEC	
Confiance	DPC	
GN1- reconnaissance à la forme verbale - []- clients	DEC	
fiabilité - PRP - [DET :POS] - produits	DEC	FIABILITE
exigence	DPC	SUIVI DES EXIGENCES
délais	DPC	MAITRISE DES DELAIS
qualité - [PREP] - [DET] - NOM	DEC	
performance - [] - produits	DEC	QUALITE DES PRODUITS
Qualification	DPC	QUALIFICATION DES PRODUITS
Qualifié	DPC	
Agréés	DPC	
Essais	DPC	
Tests	DPC	
Caractéristiques techniques	DPC	CARACTERISTIQUES TECHNIQUES
Haute précision	DPC	
Haute vitesse	DPC	
innover - PREP - GN	DEC	INNOVER
innover - PRP - verbe d'action - GN	DEC	
optimisation - Nom	DEC	
investissements sur les procédés - PRP - Nom	DEC	INVESTIR
Acquisition de procédés	DPC	
Amélioration-optimisation des procédés processus	DEC	

PATRONS	USAGE	CLASSE A ACTIVER
Acteur - [] - dans - industrie - GN	EICRC	
Partenariat - PREP - GN	EICRC	PARTENARIAT CLE
ingénierie	DPC	PROCEDES D INGENIERIE
conception - GN	EICRC	
bureau d'étude	DPC	
bureau des méthodes	DPC	
CAO	DPC	USAGE DE PROCEDES CAO
CAD	DPC	
Prototypage	DPC	USAGE DE PROCEDES DE PROTOTYPAGE
Prototype	DPC	
Outillage de contrôle	DPC	OUTILLAGE DE CONTROLE
Moyens-outil-outillage-technologie de mesure	DEC	
Marqueur - GN	DEC	
Maitrise, gestion de la qualité	DPC	DEMARCHE QUALITE
Exigence qualité	DPC	
Assurance qualité	DPC	
certification - GN	EICRC	CERTIFICATION
Représentant Entreprise - verbe d'autoréférence - certifié - GN	EICRC	
norme - GN	EICRC	NORME
normalisé-GN	EICRC	
ISO	DPC	
habilité	DPC	HABILITATION
agrée	DPC	AGREMENTS

Bibliographie

[1] IEEE 1996. *IEEE Standard for Developing Software Life Cycle Processes.* IEEE Computer Society. New York (USA), 1996. 40

[2] H. Afsarmanesh and L.M. Camarinha-Matos. A framework for management of virtual organization breeding environment. In *PRO-VE'05*, 2005. 5

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large data bases. In *J B Bocca M Jake and C Zaniolo , editors Procceding of 20 th International Conference*, 1994. 30

[4] C.H. Amherdt. Condition d'émergence des compétences collectives. aspects théoriques et étude de cas. In *4ème journée d'étude sur la Gestion des Compétences et des Connaissances en Génie Industriel*, 23 Novembre 2000 Saint Etienne. 103

[5] H. Assadi. *Construction d'ontologies à partir de textes techniques. Application aux systèmes documentaires.* PhD thesis, Université Paris 6, 1998. 38

[6] N. Aussenac-Gilles, S. Després, and S. Szulman. *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text.* IOS Press, 2008. 52

[7] B. *Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances.* Paris : L'Harmattan, 2000. 43

[8] Hu B., Y. Kalfoglou, H. Alani, P. Lewis D. Dupplaw, and N. Shadbolt. Semantic metrics. In *In Proc. 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW), Volume 4248 of Lecture notes in computer science, Praha (CZ), pp.166-181*, 2006. 153

[9] B. Bachimont, A. Isaac, and R. Troncy. Semantic commitment for designing ontologies. *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, LNAI 2473 :114–121, 2002. 36, 43, 115, 117, 118, 127

[10] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press / Addison-Wesley, 1999. 13

[11] Gilles Balmisse. *La recherche d'information en entreprise.* Lavoisier, 2007. 2

[12] R.J Bayardo, R. Agrawal, and D. Gunoplos. Constraint-based rule mining in large dense databases. In *Proceedings of the 15th Internationnal Conference on Data Engineering*, 1999. 30, 31

[13] R.K. Belew. Finding out about : A cognitive perspective on search engine technology and the www. *Review published in Information Retrieval*, 5 :Issue 2–3, April-July 2002. New York : Cambridge University Press. 13

[14] F. Belkadi, E. Bonjour, and M. Dulmet. Démarche de modélisation d'une situation de conception collaborative. *Revue Document numérique*, Vol.8 :93–106, 2004. 104

[15] M. Benali. *Une modélisation des liens de coopération et des trajectoires d'évolution des réseaux d'entreprises*. PhD thesis, L'Ecole Nationale Supérieure des Mines de Saint-Etienne et de l'Université Jean Monnet, Saint-Etienne, France, 2005. 3, 5, 6, 57, 64, 66, 149, 155, 156, 157

[16] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. In *Leen, T. K., Dietterich, T. G., and Tresp, V., editors, Advances in Neural Information Processing Systems 13, pages 932938. MIT Press*, 2001. 84

[17] G. Berio and M. Harzallah. Towards an integrating architecture for competence management. *In Special Issue "Competence Management in Industrial Processes", guest editors X.Boucher, E. Bonjour, N.Matta, Computers in Industry*, V58 issue 2, 2007. 120, 123

[18] A. Berson, S. Smith, and K. Thearling. An overview of data mining techniques : Building data mining applications for crm. *McGraw-Hill, New York*, page 488, 1999. 32

[19] E. Blanchard and M. Harzallah. Reasoning on competencies. In *In Proceedings of the Workshop on Knowledge Management and Organizational Memories (joint with ECAI2004) Valencia, Spain*, 2004. 106

[20] O. Bodenreider. The unified medical lange system (umls) : integrating biomedical terminology. In *Nucleic Acids Research, 32, Database issue :D267-70*, 2004. 36

[21] E. Bonjour and M. Dulmet. Articulation entre pilotage des systèmes de compétences et gestion des connaissances. In *1er colloque de gestion des compétences et des connaissances en génie industriel*, pages 43–50, Nantes, décembre 2002. 103

[22] G. Bordogna and G. Pasi. Flexible querying of structured documents. In *Flexible Query Answering Systems (FQAS) pages 350-361, Warsaw, Poland*, 2000. 18, 103

[23] P. Borst, H. Akkermans, and J. Top. Engineering ontologies. *J. Hum.-Comput. Stud*, 46(2) :365–406, 1997. 35

[24] X. Boucher, E. Bonjour, and B. Garabot. Formalization and use of competencies for industrial performance optimisation : a survey. *Special Issue " Competence management in Industrial Process" , guest editors X.Boucher, E.Bonjour, N.Matta, Computers in industry*, V58, issue2, 2007. 105, 106

[25] X. Boucher and P. Burlat. Vers l'intégration des compétences dans le pilotage des performances de l'entreprise. *Journal Européen des Systèmes Automatisés (JESA)*, vol. 37, N° 3 :363–390, 2003. 103, 104, 105, 121

[26] M. Boughanem and C. Soulé-Dupuy. A connexionist model for information retrieval. *DEXA*, pages 260–265, 1992. 20, 21, 84, 89, 92

[27] D. Bourigault and C. Fabre. Approche linguistique pour lanalyse syntaxique de corpus. In *Cahiers de Grammaires 25,131151*, 2000. 38

[28] D. Bourigault, C. Fabre, C. Frérot, M.-P. Jacques, and S. Ozdowska. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles, Dourdan, France.*, 2005. 52

[29] D. Bourigault and C. Jacquemin. Construction de ressources terminologiques. *J-M Pierrel (éd), Industrie des langues, Hermès Paris*, pages 215–233, 2000. 37

[30] A. Bérard-Dugourd, J.Farges, M.-C Landau, and J.-P Rogala. Natural language analysis using conceptual graphs. In *Procedeedings of the international Computer Science Conference 88, Hong-Kong, pages 265-272*, 1988. 24

[31] P. Burlat and M. Benali. A metodology to characterise co-operation links for networks of firms. *Production Planning and Control*, Vol. 18 No. 2 :156–168, March 2007. 3, 6, 57, 96, 97, 105, 106

[32] P. Burlat, D. Villa, B. Besombes, and V. Deslandres. Un cadre d'analyse dynamique des réseaux d'entreprises. *Revue Française de Gestion Industrielle (RFGI)*, 22 :77–94, 2003. 156

[33] L. Hasler C. Orasan, R. Mitkov. Cast : a computer-aided summarisation tool. In *Proceedings of EACL2003*, pages 135 – 138, Budapest, Hungary, April 2003. 27

[34] P.James Callan, W. Bruce Croft, and M.Stephen Harding. The inquery retrieval system. *DEXA*, pages 78–83, 1992. 22

[35] L. M. Camarinha-Matos, P. Macedo, and A. Abreu. Analysis of core-values alignment in collaborative networks. In *Virtual Enterprises and Collaborative Networks*, pages 53–64, 2008. 105

[36] L. M. Camarinha-Matos and W. Picard, editors. *Pervasive Collaborative Networks, IFIP TC 5 WG 5.5 Ninth Working Conference on Virtual Enterprises, September 8-10, 2008, Poznan, Poland*, volume 283 of *IFIP*. Springer, 2008. 105

[37] LM. Camarinha-Matos and H. Afsarmanesh. Elements of a base ve infrastructure. *Computers in Industry*, 51 :139–163, 2003. 4, 63, 64, 106

[38] LM. Camarinha-Matos and H. Afsarmanesh. A comprehensive modeling framework for collaborative networked organizations. *Journal of Intelligent Manufacturing*, 18 :529–542, 2007. 4, 106

[39] L.M. Camarinha-Matos, H. Afsarmanesh, and M. Ollus. *Virtual Organizations : systems and practices.* Springer Science, 2005. 5

[40] M. Chagnoux, N. Hernandez, and N. Aussenac. From text to ontologies : Non-taxonomical relation extraction. In *JFO , Lyon-France*, 2008. 37, 39, 53

[41] P. Cimiano. *Ontology Learning and Population from Text. Algorithms, evaluation and applications.* Springer, Berlin, 2007. 37, 39

[42] P. Cimiano and J. Volker. Text2onto - a framework for ontology learning and data-driven change discovery. In *the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of Lecture Notes in Computer Science, p. 227-238, Alicante, Spain : Springer.*, 2005. 52

[43] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the semantic web with the corese search engine. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004)*, 2004. 106

[44] T. Cover and P. Hart. Nearest neighbor pattern classification. information theory. *IEEE Transactions*, 13(1) :21–27, 1967. 31

[45] J.A. Crispim and J. Pinfo de Sousa. Multiple criteria partner selection in virtual enterprises. In *Proceedings of PROVE'07 8th IFIP Working Conference*, 2007. 4

[46] Z. W. Ras H. Hacid (ed.) D. A. Zighed, S. Tsumoto. *Mining Complex Data.* Springer, 2009, Vol. 165. 26

[47] B. Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* PhD thesis, Université de Paris 7, Paris, 1994. 38

[48] S. David and P. Plante. De la nécéssité dune approche morpho syntaxique dans lanalyse de textes. *In Intelligence Artificielle et Sciences Cognitives au Quebec*, 3 :140154, 1990. 38

[49] Scott Deerwester, Susan Dumais, Thomas Landauer, George Furnas, and Richard Harshman. Indexing by latent semantic analysis. *Journal American Society of Information Science*, 41 :6 :391–407, 1990. 15

[50] Lisa Di-Jorio, Lylia Abrouk, Céline Fiot, Danièle Hérin, and Maguelonne Teisseire. Enrichissement d'ontologie basé sur les motifs séquentiels. In *Plateforme AFIA 2007, Atelier Ontologies et gestion de l'hétérogénéité sémantique (OGHS)*, 2007. 37, 39

[51] M. Diana, R. Koeling, J. Weeds, and J. Carroll. Finding predominant senses in untagged text. In *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain*, 2004. 153

[52] M. Ehrig, P. Haase, M. Hefke, and N. Stojanovic. Similarity for ontology-a comprehensive framework. In *In Workshop Enterprise Modelling and Ontology : Ingredients for Interoperability*, 2004. 151

[53] Mariana Enusi. Competence management and the competence management information systems of the small and medium entreprises. *Fascicle of Management and Technological Engineering*, Volume VII (XVII), 2008. 105

[54] E. Ermilova and H. Afsarmanesh. Modeling and management of profiles and competencies in vbes. *Journal of Intelligent Manufacturing*, 18 :561–586, 2007. 105, 122, 123

[55] E. Ermilova, N. Galeano, and H. Afsarmanesh. Ecolead deliverable d21.2a. In *Specification of the VBE competency/profile management*, 2005. 4, 63, 64, 106

[56] J.A. Farrel and A.N. Michel. Associative memory via artificial neural networks. In *IEEE control systems magazine*, 1990. 85

[57] D. Faure. *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. PhD thesis, Université de Paris Sud, 2000. 38

[58] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 1 :37–54, 1996. 26

[59] L. Franchini, E. Caillaud, P. Nguyen, and G. Lacoste. Workload control of human resources to improve production management. *I. J. of Production Research*, Vol. 39 :1385–1403, 2001. 105

[60] J.M Frayret, P. Forget, and S. Amours. Un agent à comportements multiples pour la planification de la chaîne d'approvisionnement : une application à l'industrie forestière. In *7e Congrés international de génie industriel- 5-8 juin - Trois Rivières Québec*, 2007. 4

[61] F. Fürst. *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation*. PhD thesis, Université de Nantes, France, Novembre, 2004. 39, 116

[62] F. Gandon. Ontology engeneering : a survey and a return on experience. In *rapport de recherche 4396,INRIA,.*, 2002. 45

[63] M. Gluck and D. Rumelhart. *Neuroscience and connectionist theory*. Lawrence Erlbaum, London., 1990. 85

[64] P.A. Gomez and D. Rojas-Amaya. Ontological reengineering for reuse. *11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW-99) LNAI, Berlin*, 1621 :139–156, 1999. 39, 40, 116

[65] N. Grabar and T. Hamon. Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'intelligence artificielle*, 18(1) :57–85, 2004. 52

[66] B. Grabot and A. Letouzy. Short-term manpower management in manufacturing systems : new requirement and dss prototyping. *Computers in Industry*, Vol.43 :11–29, 2000. 105

[67] T.R. Grüber. Toward principles for the design of ontologies used for knowledge sharing. In *International Workshop on Formal Ontology, Padova, Italy, March*, 1993. 44

[68] M. Grundstein. De la capitalisation des connaissances au renforcement des compétences dans l'entreprise étendue. In *Conférence invitée, 1er colloque du groupe de travail "Gestion des Compétences et des Connaissances en Génie Industriel", Nantes*, 2002. 103

[69] David Gsman and Ophir Frieder. *Ad Hoc Information Retrieval : Algorithms and Heuristics*. Kluwer Academic Publishers, 1998. 13

[70] N. Guarino. Semantic matching : Formal ontological distinctions for information organization, extraction, and integration. *SCIE*, pages 139–170, 1997. 35, 44

[71] N. Guarino, C. Masolo, and G. Vetere. Ontoseek : Using large linguistic ontologies for accessing on-line yellow pages and product catalogs. In *National Research Council, LADSEBCNR : Padavo, Italy*, 1999. 76

[72] N. Guarino and L. Schneider. Ontology-driven conceptual modelling. In *ER*, page 10, 2002. 39, 116

[73] G. Ritschard D. A. Zighed (ed.) H. Briand, F. Guillet. *Advances in Knowledge Discovery and Management.* Springer, 2009. 26

[74] B. Habert, E. Naulleau, and A. Nazarenko. Symbolic word clustering for medium-size corpora. In *the 16th International Conference on Computational Linguistics (CoLing'96), Copenhagen, pp 490-495*, 1996. 38

[75] M. Haddad. *Extraction et impact des connaissances sur les performances des systèmes de recherche d'information.* PhD thesis, l'Université de Grenoble 1, 2002. 52

[76] K. Hajlaoui. Information extraction procedure to support the constitution of virtual organisations. In *IEEE International Conference on Research Challenges in Information Science (RCIS 2008) Marrakech, Morocco*, 2008. 73

[77] K. Hajlaoui and X. Boucher. Neural network based text mining to discover enterprise networks. In *13th IFAC Symposium on Information Control Problems in Manufacturing (INCOM'2009). Moscow, Russia*, 2009. 86

[78] K. Hajlaoui, X. Boucher, and J.J Girardot. Competency ontology for network building. In *10th IFIP Working Conference on Virtual Entreprises (PRO-VE'09). Thessaloniki, GREECE*, 2009. 119, 123

[79] K. Hajlaoui, X. Boucher, and M. Mathieu. Data mining for the identification of virtual organisations. In *9th IFIP Working Conference on Virtual Enterprises (PRO-VE'08). Poznan, POLAND*, 2008. 73

[80] M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *In A. Zampolli, editor, Computational Linguistics (CoLing'1992), pages 539-545, Nantes, France,*, 1992. 39, 52

[81] William R. Hersh, Chris Buckley, T. J. Leone, and David H. Hickam. Ohsumed : An interactive retrieval evaluation and new large test collection for research. In *SIGIR 192-201*, 1994. 76

[82] G. Hirst and A. Budanitsky. Correcting real-word spelling errors by restoring lexical cohesion. In *Natural Language Engineering*, 2004. 153

[83] J. Hodík, J. Vokrínek, J. Bíba, and P. Becvár. Competencies and profiles management for virtual organizations creation. In *CEEMAS*, 2007. 121, 123

[84] I. Horvath and J.J Broek. Advanced computer support of engineering and service processes of virtual enterprises. *Editorial, Special Issue, Computers in industry*, 57 :201–203, 2006. 4, 106

[85] F. Ibekwe-Sanjuan. *Fouille de textes : méthodes, outils et applications.* éditions Hermès-Lavoisier, 2007. 28

[86] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *In Proceedings of International Conference on Research in Computational Linguistics, Taiwan*, 1997. 151

[87] T. Joachims. Text categorization with support vector machines : learning with many relevant features. In *10th European Conference on Machine Learning ECML-98, pp.137-142.*, 1998. 31

[88] K.S Jones and P. Willett. *Readings in Information Retrieval.* Morgan Kaufmann Publishers, 1997. 15

[89] Sparck Jones K. Experiments in relevance weighting of search terms. *Information Processing and Management*, 15(3) :133–144, 1979. 15

[90] H. Kamp. *A theory of truth and semantics representation.* In Groenendijk, Jansen and Stokhof, Eds., Formal Methods in the Study of Language. Amsterdam : Mathematical Centre Tracts, 1981. 51

[91] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis.* Wiley Interscience, New York, 1990. 28

[92] M. Keller and S. Bengio. A neural network for text representation. In *In Proceedings of the 15th International Conference on Artificial Neural Networks : Biological Inspirations, ICANN, Lecture Notes in Computer Science, volume LNCS 3697, pages 667672*, 2005. 84

[93] L. R. Khan. *Ontology-based Information Selection.* PhD thesis, Faculty of the Graduate School, University of Southern California, 2000. 76

[94] T.Y. Kim, S. Lee, K. Kim, and C.H Kim. A modeling framework for agile and interoperable virtual enterprises, in advanced computer support of engineering and service processes of virtual enterprises. *Special Issue, Computers in industry*, 57 :201–203, 2006. 4, 106

[95] A. Kontostathis, L.M. Galitsky, W.M. Pottenger, Soma Roy, and Daniel J. Phelps. *A Survey of Emerging Trend Detection in Textual Data Mining*. Springer, 2004. 27

[96] J. Kuhn and L. Maron. On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery*, 7(3) :216–244, 1960. 21

[97] K.L. Kwok. A neural network for probabilistic information retrieval. In *12th International ACM SIGIR Conference on Research and Developpement in Information Retrieval, pp 21-30*, 1989. 20, 92

[98] K.L. Kwok. A network approach to probabilistic information retrieval. In *ACM transactions on information systems. Pages 324-353*, 1995. 20

[99] G. Lame. *Construction d'ontologie à partir de textes. Une ontologie du Droit français dédiée à la recherche d'information sur le Web*. PhD thesis, Ecole des Mines de Paris, 2002. 38

[100] T.K. Landauer and S.T. Dumais. A solution to plato's problem : the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104 :211–240, 1997. 30

[101] M. Laukkanen and H. Helin. Competence management within and between organizations. In *EMOI-INTEROP*, 2005. 106

[102] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In *In WordNet : An Electronic Lexical Database, C. Fellbaum, MIT Press*, 1998. 152

[103] J.H Lee, M.H Kim, and Y.J Lee. Information retrieval based on conceptual distance in is-a hierarchy. *Journal of Documentation*, 49 :188–207, 1993. 151

[104] Yeong Su Lee and Michaela Geierhos. Business specific online information extraction from german websites. In *CICLing '09 : Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 369–381, Berlin, Heidelberg, 2009. Springer-Verlag. 63

[105] J. Leplat. A propos des compétences. *Revue EPS*, pages 267, 9–12, 1997. 103

[106] D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1 : A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5 :361–397, Apr 2004. 31

[107] D. Lin. An information-theoretic definition of similarity. In *In Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98) Morgan- Kaufmann : Madison, WI*, 1998. 151

[108] X. Lin, D. Soergel, and G. Marchionini. A self organizing semantic map for information retrieval. In *SIGIR 91, Chicago, Illinois*, 1991. 21

[109] Y. Loiseau, M. Boughanem, and H. Prade. Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In *AMR05, 3rd International Workshop on Adaptive Multimedia Retrieval, Glasgow, UK*, 2005. 18

[110] J.B. MACQUEEN. Some methods for classication and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability,Berkeley, CA, University of California Press, no 1*, page 281.297, 1996. 29

[111] W. McCulloch and W. Pitts. *A logical calculus of ideas immanent in nervous activity.* Bulletin of Mathematical Biophysics, 1943. 84, 85

[112] D.L. Medin. Concepts and conceptual structure. *American Psychologist*, 44(12) :1469–1481, 1989. 36

[113] O. Mendes. *État de l'art sur les méthodologies d'ingénierie ontologique.* PhD thesis, Montréal, Québec, Canada : Centre de recherche LICEF, 2003. 40

[114] F.X Micheloud. *L'analyse des correspondances.* PhD thesis, Ecole des hautes Etudes Commerciales, publié sur le site http ://www.micheloud.com/FXM/COR/index.htm, Lausanne, 1997. 29

[115] P. Milgrom and J. Roberts. *Economie, organisation et management.* Université, Bruxelles, Belgique, 1997. 66, 96

[116] G. Miller. Wordnet : A lexical database. In *Communication of the ACM, 38(11) :39–41*, 1995. 36

[117] R. Mizoguchi. A step towards ontological engineering. In *the 12th National Conference on AI of JSAI*, 1998. 40

[118] T. Mondary, S. Despres, A. Nazarenko, and S. Szulman. Construction d'ontologies à partir de textes : la phase de conceptualisation. In *Construction d'ontologies à partir de textes : la phase de conceptualisation*, 2008. 52

[119] J. Mothe. *Modèle Connexionniste pour la Recherche d'Information, Expansion dirigée de requêtes et apprentissage.* PhD thesis, l'Université Paul Sabatier, Toulouse (France), 1994. 20, 84, 89, 92

[120] M.C. Mozer. Inductive information retrieval using parallel distributed computation. In *Institute for Cognitive Science (ICS) T.R. 84 06. La Jolla : UCSD*, 1984. 21

[121] N. Ben Mustapha, R. Soussi, H.B. Zgal, and M. Aufre. A metaontology for domain ontology enriching in an information retrieval system. In *JFO (Journées Francophones sur les Ontologies) 2008 Lyon-France*, 2008. 53

[122] M.Mc Cord Nelson and W.T Illingworth. *A practical guide to neural nets.* Addison Wesely, 1990. 85

[123] A. Opdahl and G. Berio. *A Roadmap for UEML, Enterprise interoperability :New challenges and approaches.* Springer edition. ISBN : 978-1-84628-713-8, p.189-198, 2006. 2

[124] H. Panetto. *Meta-Modèles et Modèles pour l'Intégration et l'Interopérabilité des Applications d'Entreprises de Production.* PhD thesis, HDR, Université Nancy 1, 2006. 2

[125] V. Parekh, J.P. Gwo, and T. Finin. Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In *International Conference of Information and Knowledge Engineering*, 2004. 37

[126] S. Paumier. *Recherche d'expressions dans de grands corpus : le système AGLAE.* PhD thesis, Master thesis,Université de Marne-la-Vallée., 2000. 54

[127] S. Paumier. *De La reconnaissance de formes linquistique a l'analyse syntaxique.* PhD thesis, Marne-la-Vallèe, 2003. 54

[128] S. Paumier. *Unitex 1.2 Manuel d'utilisation.* Universite Marne-la-Vallèe, 2004. 54

[129] S. Paumier. *Unitex 1.2 Manuel d'utilisation.* Universite Marne-la-Vallèe, 2006. 54

[130] S. Peillon. *Le pilotage des coopérations inter-entreprises : le cas des groupements de PME.* PhD thesis, l'Université Jean Monnet, 2001. 64, 156

[131] J. Plisson, P. Ljubic, I. Mozetic, and N. Lavrac. An ontology for virtual organisation breeding environments. In *To appear in IEEE Trans. On Systems, Man, and Cybernetics*, 2007. 4, 63, 64, 106

[132] G. Pépiot. *Modélisation des Entreprises sur la base des compétences.* PhD thesis, EPFL, 2005. 120, 123

[133] V. Psyché, R. Mizoguchi, and B. Bourdeau. Ontology development at the conceptual level for theory-aware its authoring systems. In *Conference on Artificial Intelligence in Education (AIED03)*, 2003. 40

[134] J.R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1) :81–106, 1986. 32

[135] J.R. Quinlan. *programs for machine learning.* San Francisco, CA, USA : Morgan Kaufmann Publishers Inc, 1993. 32

[136] P. Resnik. Using information content to evaluate semantic similarity in taxonomy. In *In Proceedings of 14th International Joint Conference on Artificial Intelligence, Montreal*, 1995. 151

[137] P. Resnik. Semantic similarity in a taxonomy : An information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11 :95–130, 1999. 152

[138] G.B. Richardson. the organization of industry. *Economic Journal*, vol.82 :883, 1972. 66

[139] C.J Van Rijsbergen. *Information retrieval.* London : Butterworth,, 1979. 13, 22

[140] E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at trec. *TREC*, pages 21–30, 1992. 22

[141] S.E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33 (4) :294–304, 1977. 21

[142] S.E. Robertson and S. K. Jones. Relevance weighting of search terms. *Journal American Society for Information Science*, 27 :129–146, 1976. 21, 22

[143] R. Studer S. Staab, H.P. Schnurr, and Y. Sure. Knowledge processes and ontologies. *IEEE Intelligent Systems*, pages 26–34, 2001. 36, 40, 41

[144] G. Sabah. Sens et traitements automatiques des langues. *dans J.-M Pierrel (dir.)*, pages 77–129, 2001. 24

[145] G. Salton. Search and retrieval experiments in real-time information retrieval. In *International Federation for Information Processing (IFIP) Congress (2) : 1082-1093*, 1968. 18

[146] G. Salton. A comparison between manual and automatic indexing methods. *Journal of the American Documentation*, 20(1) :6171, 1971. 15

[147] G. Salton, E.A. Fox, and H. Wu. Extended boolean information retrieval system. *Communications of the ACM*, 26(11) :1022–1036, 1983. 18, 24, 76

[148] H. Samelides, P. Bouret, and J. Reggia. *Réseaux neuronaux une approche connexionniste de l'intelligence artificielle.* édition TEKNEA, 1991. 85

[149] E. SanJuan and F. Ibekwe-SanJuan. Textmining without document context. *Information Processing and Management, Special issue on Informetrics II Elsevier*, 42(6) :1532–1552, 2006. 28

[150] A. Schutz and P.Buitelaar. Relext : A tool for relation extraction from text in ontology extension. In *In Y. G. et al., editor, ISWC 2005, LNCS 3729, pages 593-606*, 2005. 37, 84

[151] M. Silberztein. *Dictionnaires électronique et analyse automatique de texte, le systeme INTEX.* Masson, 1993. 53, 54

[152] JK. Startman. Realizing benefits from enterprise resource planning : Does strategic focus matter. *Production and Operations Management*, Vol. 16 No. 2 :203216, 2007. 105

[153] G. Stumme, A. Hotho, and B. Berendt. Semantic web mining : State of the art and future directions. *Web Semantics : Science, Services and Agents on the World Wide Web*, 4(2) :124–143, 2006. 37, 39

[154] Y. Sure, A. Maedche, and S. Staab. Leveraging corporate skill knowledge : from proper to ontoproper. In *Processings of the 3rd Internationnal Conference on Practical Aspects of knowledge Management, Basel, Switzerland*, 2000. 106

[155] B. Swartout, R. Patil, K. Knight, and T. Russ. Towards distributed use of large-scale ontologies. In *Spring Symposium Series on Ontological Engineering, pp.138-148*, 1997. 42

[156] Jihed Touzi. *Aide à la conception de Système d'Information Collaboratif support de l'interopérabilité des entreprises.* PhD thesis, Institut National Polytechnique de Toulouse, 2007. 3

[157] M. Uschold and M. King. Towards a methodology for building ontologies. In *Basic Ontological Issues in Knowledge Sharing, Inter. Conf. on Artificial Intelligence (IJCAI)*, 1995. 41

[158] D. Vanderhaegen and P. Loos. Distributed model management platform for cross-enterprise business process management in virtual enterprise networks. *Journal of Intelligent Manufacturing*, 18 :553–559, 2007. 4, 106

[159] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, N.Y, 1995. 31

[160] P. Velardi, M. Missikoff, and R. Basili. Identification of relevant terms to support the construction of domain ontologies. In *ACL WS on Human Language Technologies and Knowledge Management. Toulouse-France*, 2001. 38

[161] P. Vossen. *A Multilingual Database with Lexical Semantic Networks.* Dordrecht, Kluwer, 1998. 36

[162] D. Vrandeci and Y. Sure. How to design better ontology metrics. In *In Proc. 4th European Semantic Web Conference, Innsbruck (AT), Volume 4519 of Lecture Notes in Computer Science, pp. 311-325*, 2007. 153

[163] D.A. Waterman. *A Guide to Expert Systems.* Addison-Wesley. Boston, Masachusets (USA), 1986. 40

[164] S.M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau. *Text Mining : Predictive Methods for Analyzing Unstructured Information.* Springer, 2005. 32

[165] D. Wenzek. *Construction de réseaux de neuronnes.* PhD thesis, INPG Grenoble France, 1993. 85, 86

[166] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics pp 133- 138*, 1994. 151

[167] P. Xu, A. Emami, and F. Jelinek. Training connectionist models for the structured language model. In *In Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2003. 84

[168] Y. Yang. A study of thresholding strategies for text categorization. In *Actes In Conference on Research and Development in Information Retrieval (ACM-SIGIR) , 137-145*, 2001. 31

[169] R. Yangarber. *Scenario Customization for Information Extraction.* PhD thesis, New York University, 2000. 53

[170] M. Yousfi-Monod and V. Prince. Compression de phrases par élagage de leur arbre morpho-syntaxique. *Revue des Sciences et Techniques Informatiques*, 25 :437–468, 2006. 27

[171] Y. Yussopova and A.R. Probst. Business concepts ontology for an enterprise performance and competences management. *in Special Issue " Competence Management in Industrial Processes ", guest editors X.Boucher, E. Bonjour, N.Matta, Computers in Industry*, V58, February 2007. 121, 123

[172] Ali Zaidat. *Specification d'un cadre d'ingenierie pour les reseaux d'organisations.* PhD thesis, L'Ecole Nationale Superieure des Mines de Saint-Etienne et de l'Universite Jean Monnet, 2005. 5

[173] M. Zloof. Query-by-example : A data base language. *IBM Systems Journal*, 16(4) :324–343, 1977. 20

[174] P. Zweigenbaum and N. Grabar. Liens morphologiques et structuration de terminologie. In *Actes de la conférence Ingénierie des Connaissances (IC 2000), Toulouse-France*, 2000. 24